Important formulas for DRS 111

Chapter 1

Statistical terms? Two branches of statistics

Types of data

Measurement level for each variable

Four basic sampling techniques

Difference between an observational and an experimental study

Uses and misuses of statistics

Importance of computers and calculators in statistics

Chapter 2

Formula for the percentage of values in each class: $\% = \frac{f}{n} . 100$

Where f = frequency of class, n = total number of values

Formula for the range: $R = highest\ value - lowest\ value$

Formula for the class width: $Class\ width = upper\ boundary - lower\ boundary$

Formula for the class midpoint:$X_m = \frac{lower\ boundary + upper\ boundary}{2}$ or $X_m = \frac{lower\ limit + upper\ limit}{2}$

Formula for the degrees for each session of a pie graph: $Degrees = \frac{f}{n} . 360°$

Chapter 6 Methods of Data Collection

Chapter 7 Measurement in Management and Scaling Techniques

Chapter 8 Questionnaire Design and Fieldwork Plan

Chapter 9 Sampling Design: Theory and Practice

Yangon University of Economics

PGDRS-111

Data Structure and Data Collection Methods

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Chapter 1
The Nature of Probability and Statistics

Chapter 1 Overview

1-1 Descriptive and Inferential Statistics

1-2 Variables and Types of Data

1-3 Data Collection and Sampling Techniques

1-4 Experimental Design

1-5 Computer and Calculators

Objectives

After competing this chapter, you should be able to:

- Demonstrate knowledge of statistical terms
- Differentiate between the two branches of statistics
- Identify types of data
- Identify the measurement level for each variable
- Identity the four basic sampling techniques
- Explain the difference between an observational and an experimental study
- Explain how statistics can be used and misused
- Explain the importance of computers and calculators in statistics

Statistics is the science of conducting studies to

- collect,
- organize,
- summarize,
- analyze, and
- draw conclusions from data.

1-1 Descriptive Statistics and Inferential Statistics

Statistics is divided into two main areas, depending on how data are used.

The two areas are

- Descriptive statistics
- Inferential statistics

Descriptive Statistics

Descriptive Statistics consists of the

- Collection,
  - Organization,
    - Summarization, and
      - Presentation of data.

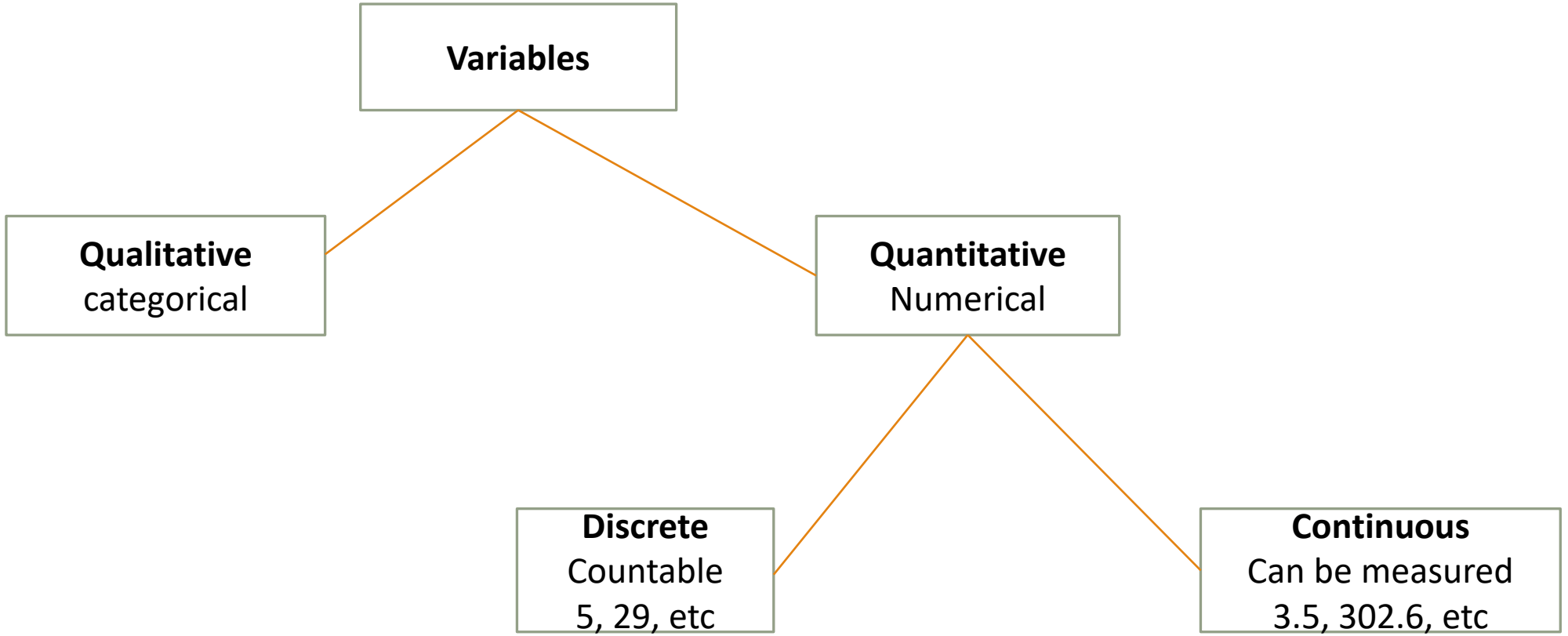
Inferential Statistics

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

- **A variable** is a characteristic or attribute that can assume different values.
- **A population** consists of all subjects (human or otherwise) that are being studied.
- **A sample** is a group of subjects selected from a population.

1-2 Variables and Types of Data

The classification of variables can be summarized as follows:



1-2 Variables and Types of Data

- **Qualitative variables** are variables that have distinct categories according to some characteristic or attribute.
- For example: blood type, voting opinion, the brand of car, Gender

- **Quantitative variables** are variables that can be counted or measured.
- For example: number of pizza sold one day, test score, temperature and income
- **Discrete variables** assume values that can be counted.
- **Continuous variables** can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

Levels of Measurement

- Nominal- categorical (names)
- Ordinal - nominal, plus can be ranked (order)
- Interval - ordinal, plus precise differences between unit of measure
- Ratio -interval, plus ratios are consistent, true zero

Examples for Levels of Measurement

Variable	Nominal	Ordinal	Interval	Ratio	Level
Hair color	Yes	No			Nominal
Pizza size	Yes	Yes			Ordinal
Height	Yes	Yes	Yes	Yes	Ratio
Age	Yes	Yes	Yes	Yes	Ratio
Temperature	Yes	Yes	Yes	No	Interval

Data Sources

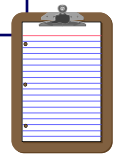
Primary

Data Collection

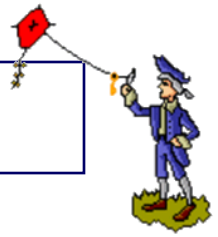


Observation

Survey



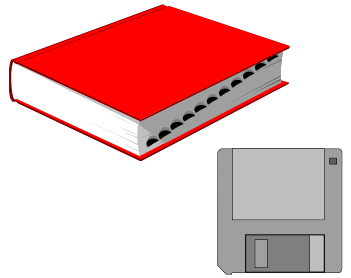
Experimentation



Secondary

Data Compilation

Print or Electronic



Population and Sample

A **Population** is the set of all items or individuals of interest

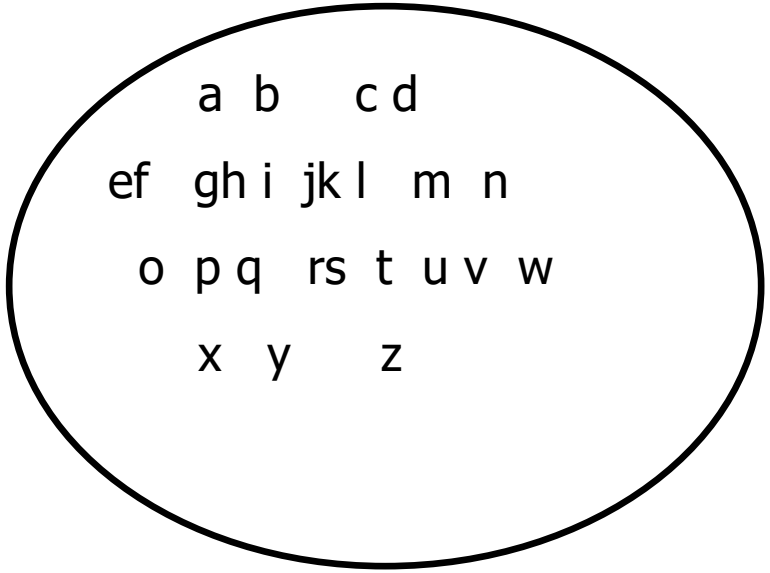
- Examples: All likely voters in the next election
  All parts produced today
  All sales receipts for November

A **Sample** is a subset of the population

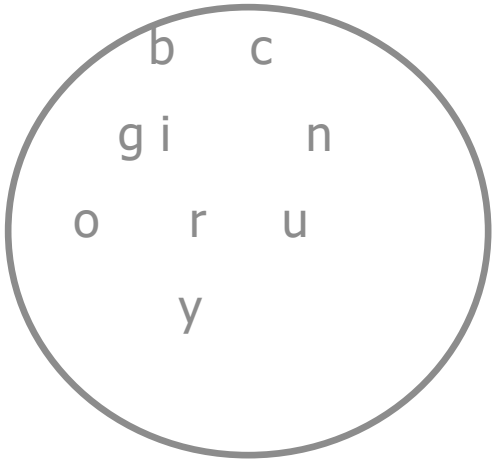
- Examples: 1000 voters selected at random for interview
  A few parts selected for destructive testing
  Every 100th receipt selected for audit

Population and Sample

Population



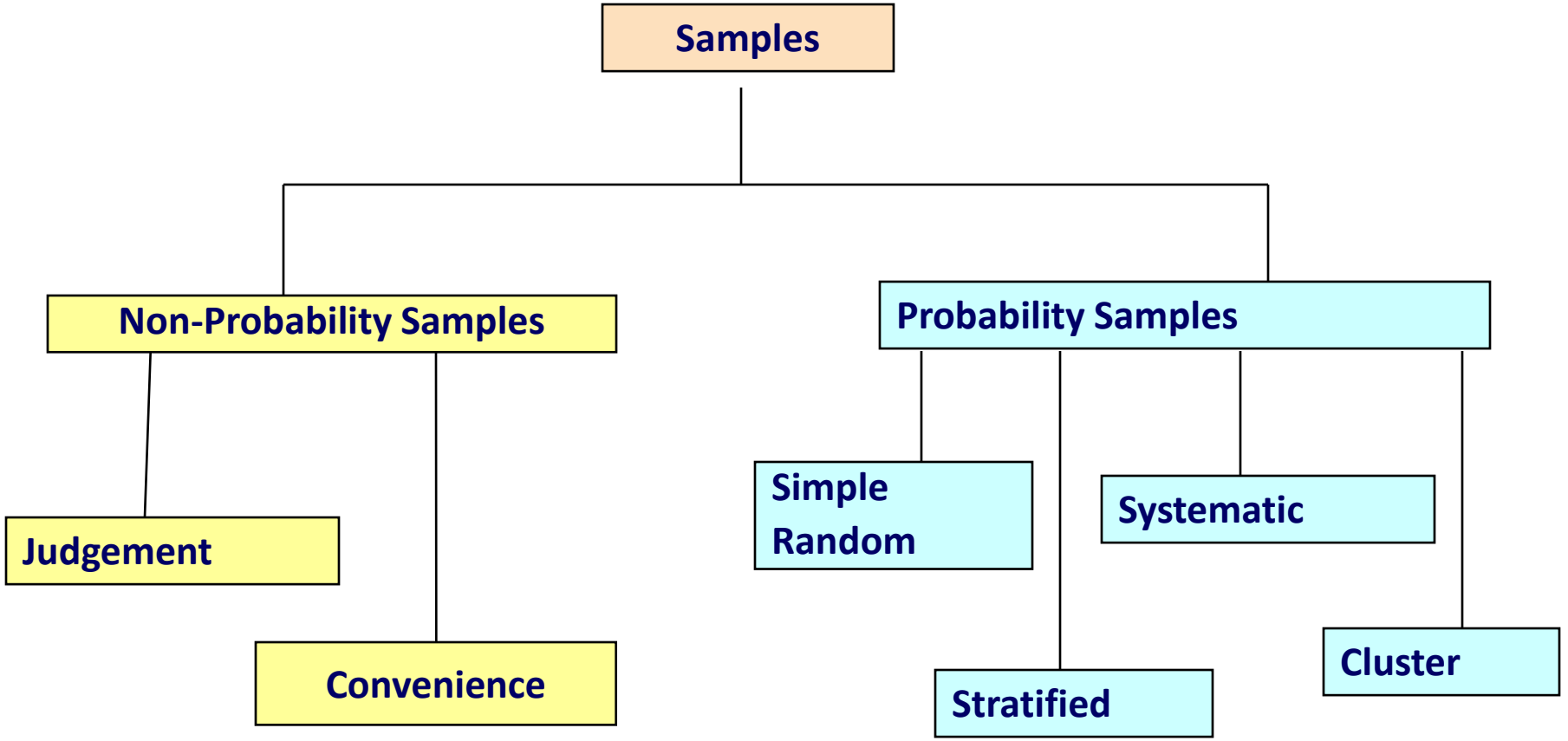
Sample



Why Sample?

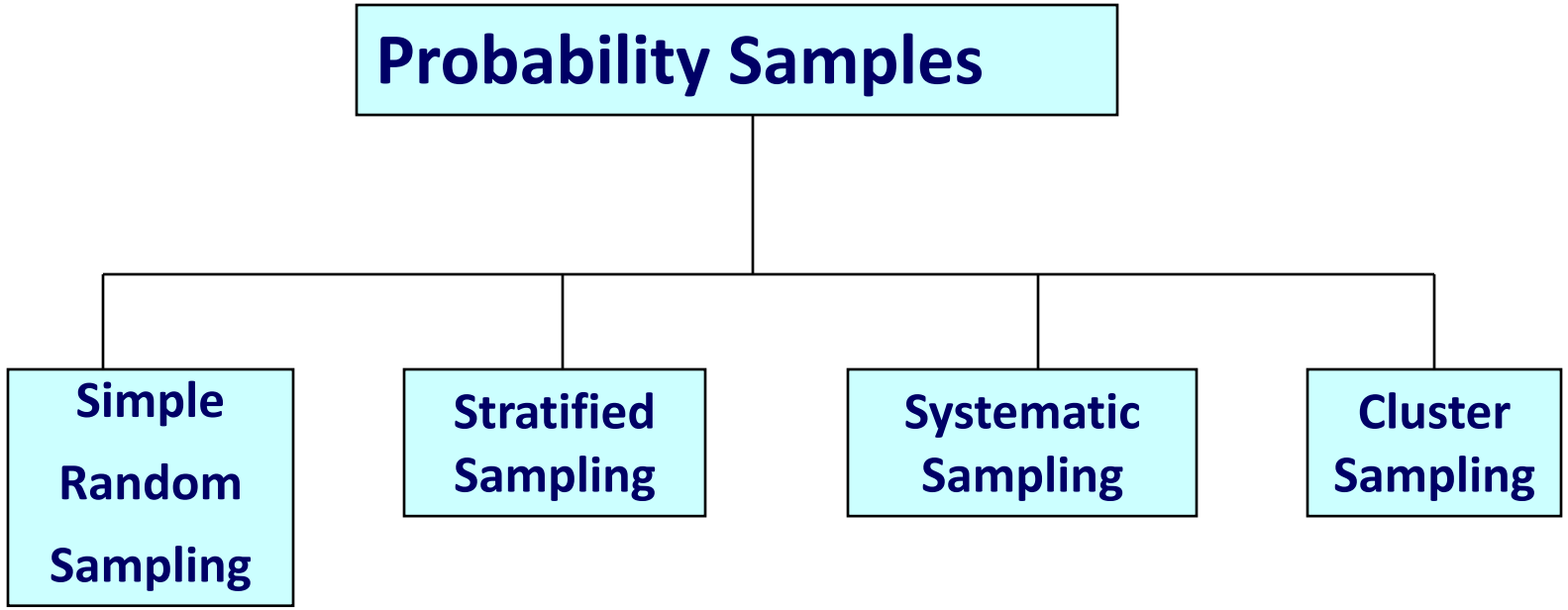
- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Sampling Techniques



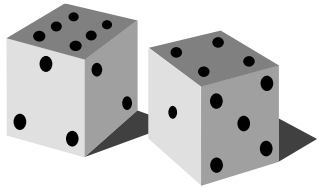
Statistical Sampling

Items of the sample are chosen based on known or calculable probabilities



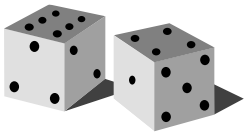
Simple Random Sample

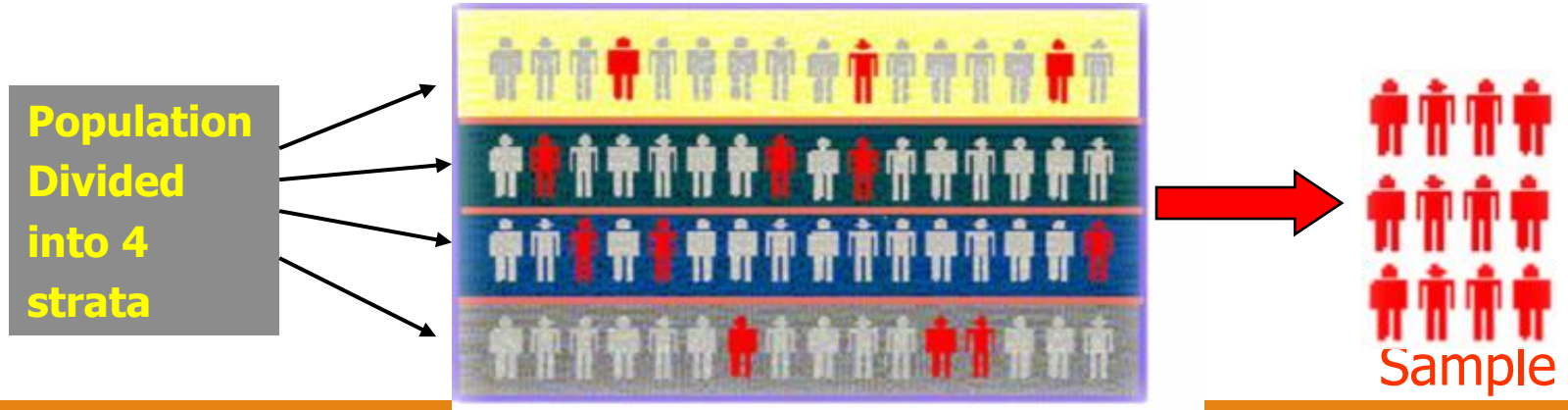
- Every individual or item from the population has an equal chance of being selected
- Selection may be with replacement or without replacement
- Samples can be obtained from a table of random numbers or computer random number generators



Stratified Sample

- Population divided into subgroups (called *strata*) according to some common characteristics
- Simple random sample selected from each subgroup
- Samples from subgroups are combined into one





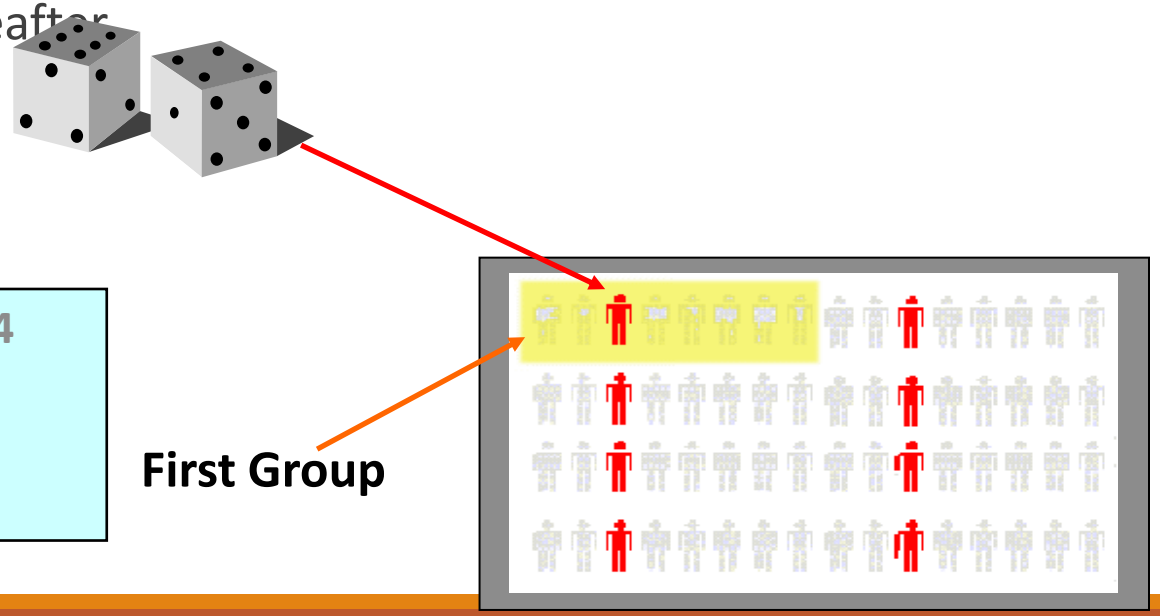
Systematic Sample

Decide on sample size: n

Divide frame of N individuals into groups of k individuals: k=N/n

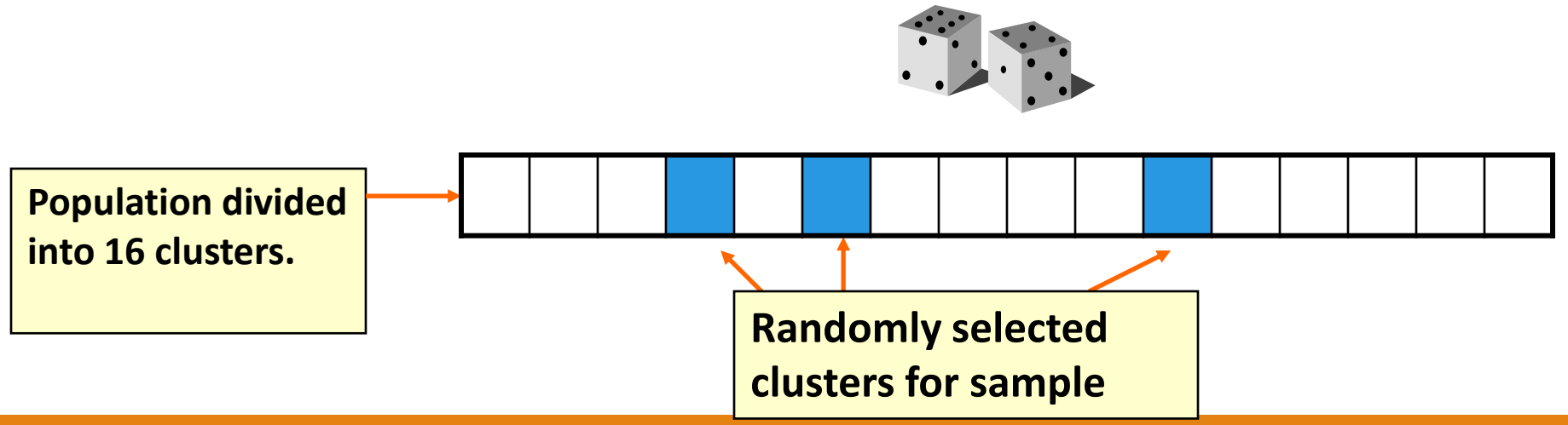
Randomly select one individual from the 1st group

Select every kth individual thereafter



Cluster Sample

- Population is divided into several clusters, each representative of the population
- A simple random sample of clusters is selected
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique



Key Definitions

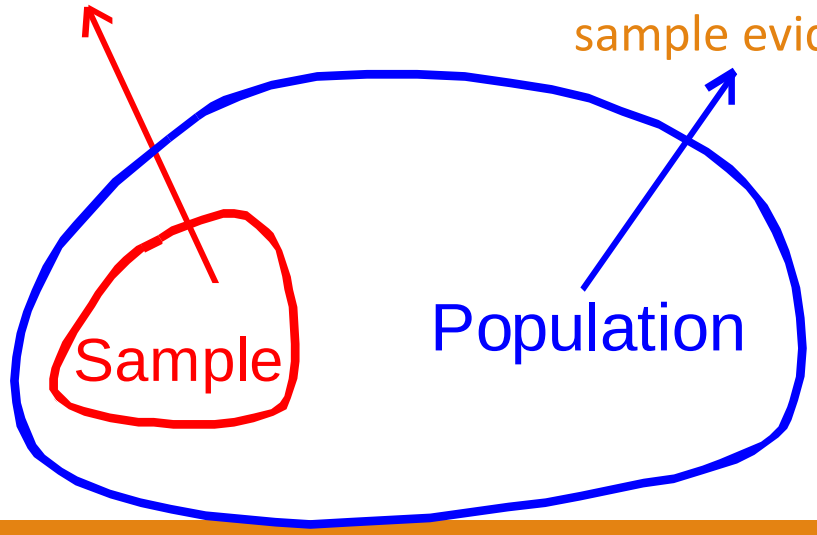
- A **population** is the entire collection of things under consideration
  - A **parameter** is a summary measure computed to describe a characteristic of the population

- A **sample** is a portion of the population selected for analysis
  - A **statistic** is a summary measure computed to describe a characteristic of the sample

Inferential Statistics

Making statements about a population by examining sample results

Sample statistics (known) **Inference** → Population parameters (unknown, but can be estimated from sample evidence)



Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based on sample results.

Estimation

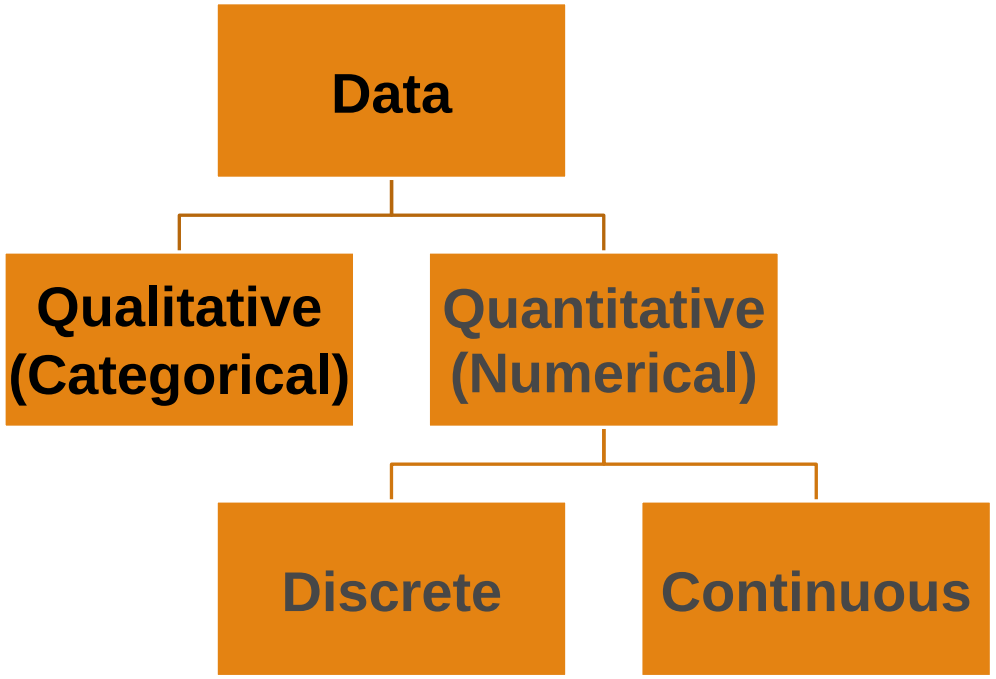
- e.g: Estimate the population mean weight using the sample mean weight

Hypothesis Testing

- e.g: Use sample evidence to test the claim that the population mean weight is 120 pounds



Data Types



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Data Types

Time Series Data

- Ordered data values observed over time

Cross Section Data

- Data values observed at a fixed point in time

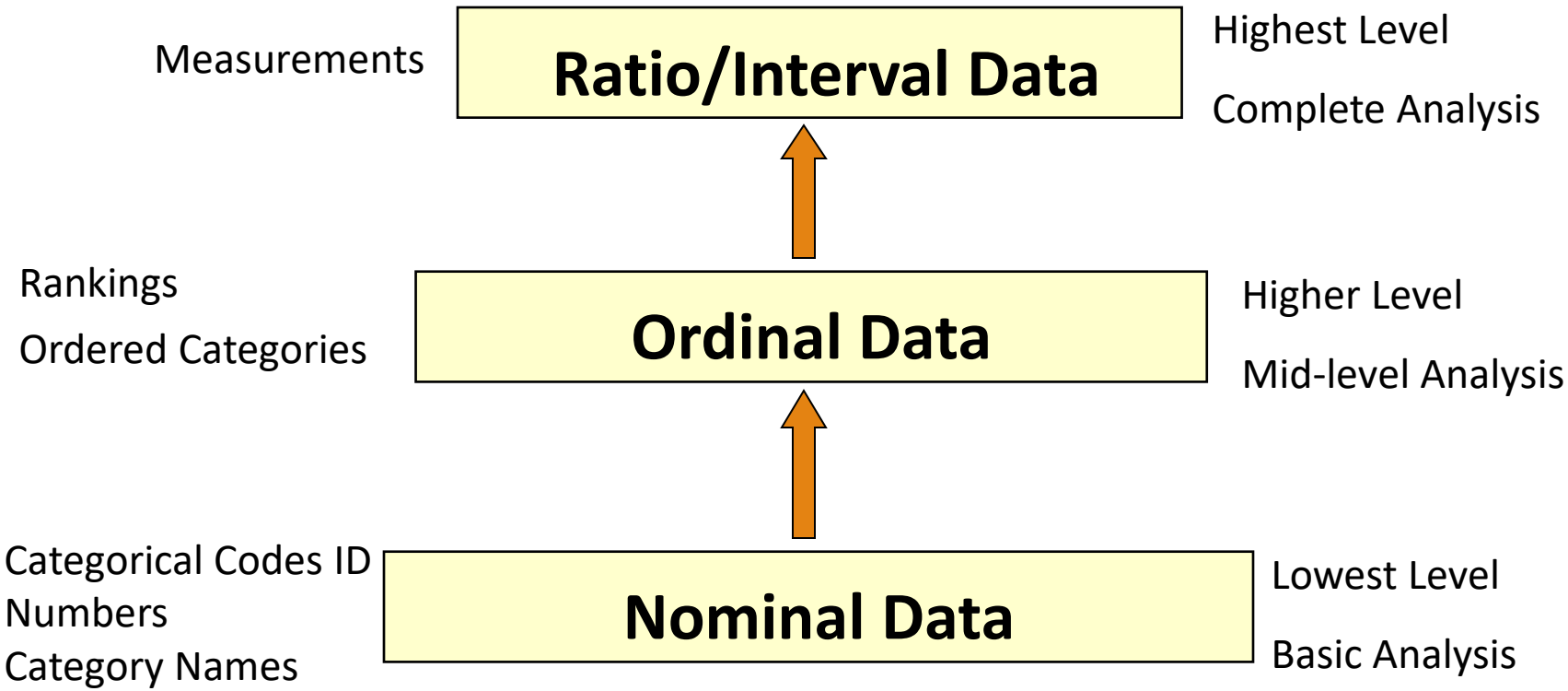
Data Types

	Sales (in $1000's)			
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Time Series Data

Cross Section Data

Data Measurement Levels



Thank You

Chapter 2

Frequency Distributions and Graphs

Dr. Hlaing Hlaing Moe

Professor/ Head

Department of Applied Statistics

Yangon University of Economics

Introduction

- 2-1 Organizing Data
- 2-2 Histograms, Frequency Polygons and Ogives
- 2-3 Other Types of Graphs
- 2-4 Paired Data and Scatter Plots

Objectives

- Organize data using frequency distributions
- Represent data in frequency distributions graphically using histograms, frequency polygons, and ogives.
- Represent data using Pareto charts, time series graphs, and pie graphs.
- Draw and interpret a stem and leaf plot.
- Draw and interpret a scatter plot for a set of paired data.

2-1 Organizing Data

- Data collected in original form is called **raw data**.
- A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.
- Nominal- or ordinal-level data that can be placed in categories is organized in **categorical frequency distributions**.

Categorical Frequency Distribution

- Twenty-five army indicates were given a blood test to determine their blood type.
- Raw Data: A,B,B,AB,O,O,O,B,AB,B,B,B,O,A,O,A,O,O,O, AB, AB,A,O,B,A

Class	Tally	Frequency	Percent
A	𝍸	5	20
B	𝍸 II	7	35
O	𝍸 IIII	9	45
AB	IIII	4	16

Grouped Frequency Distribution

- **Grouped frequency distributions** are used when the range of the data is large.
- The smallest and largest possible data values in a class are the ***lower*** and ***upper class limits***. ***Class boundaries*** separate the classes.
- To find a class boundary, average the upper class limit of one class and the lower class limit of the next class.

Grouped Frequency Distribution

- The **class width** can be calculated by subtracting
  - successive lower class limits (or boundaries)
  - successive upper class limits (or boundaries)
  - upper and lower class boundaries

- The ***class midpoint*** X_m can be calculated by averaging
  - upper and lower class limits (or boundaries)

Rules for Classes in Grouped Frequency Distributions

- There should be 5-20 classes.
- The class width should be an odd number.
- The classes must be mutually exclusive.
- The classes must be continuous.
- The classes must be exhaustive.
- The classes must be equal in width (except in open-ended distributions).
- Class width=Range/no. of classes

Constructing a Grouped Frequency Distribution

The following data represent the record high temperatures for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Range<10 ⟹ Ungrouped frequency distribution

Range>10 ⟹ Grouped frequency distribution

Class limit	Tally	Frequency
100-104		
105-109	/	1
110-114	/𝍸/	5
115-119	/	1
120-124		
125-129		
130-134		

Constructing a Grouped Frequency Distribution

STEP 1 Determine the classes.

Find the class width by dividing the range by the number of classes 7.

$$Range = High - Low$$

$$= 134 - 100 = 34$$

$$Class\ Width = Range/7 = 34/7 = 5$$

Rounding Rule: Always round up if a remainder.

Constructing a Grouped Frequency Distribution

- For convenience sake, we will choose the lowest data value, 100, for the first lower class limit.
- The subsequent lower class limits are found by adding the width to the previous lower class limits.

Class Limits
100 - 104
105 - 109
110 - 114
115 - 119
120 - 124
125 - 129
130 - 134

- The first upper class limit is one less than the next lower class limit.
- The subsequent upper class limits are found by adding the width to the previous upper class limits.

Constructing a Grouped Frequency Distribution

- The class boundary is midway between an upper class limit and a subsequent lower class limit. 104,104.5,105

Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5		
105 - 109	104.5 - 109.5		
110 - 114	109.5 - 114.5	105-104=1/2=0.5	
115 - 119	114.5 - 119.5	Lower-0.5	
120 - 124	119.5 - 124.5	Upper-0.5	
125 - 129	124.5 - 129.5		
130 - 134	129.5 - 134.5		

Constructing a Grouped Frequency Distribution

STEP 2 Tally the data.

STEP 3 Find the frequencies.

Class Limits	Class Boundaries	Frequency	Class Midpoint
100 - 104	99.5 - 104.5	2	102
105 - 109	104.5 - 109.5	8	107
110 - 114	109.5 - 114.5	18	112
115 - 119	114.5 - 119.5	13	117
120 - 124	119.5 - 124.5	7	122
125 - 129	124.5 - 129.5	1	127
130 - 134	129.5 - 134.5	1	132

Constructing a Grouped Frequency Distribution

STEP 4 Find the cumulative frequencies by keeping a running total of the frequencies.

Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5	2	2
105 - 109	104.5 - 109.5	8	10
110 - 114	109.5 - 114.5	18	28
115 - 119	114.5 - 119.5	13	41
120 - 124	119.5 - 124.5	7	48
125 - 129	124.5 - 129.5	1	49
130 - 134	129.5 - 134.5	1	50

THANK YOU

CHAPTER 2

Frequency Distributions and Graphs

Dr. Hlaing Hlaing Moe

Professor/ Head

Department of Applied Statistics

Yangon University of Economics

Chapter 2
Overview

- 2-2 Histograms, Frequency Polygons and Ogives

Chapter 2
Objectives

- Represent data in frequency distributions graphically using histograms, frequency polygons, and ogives.

Histograms, Frequency Polygons, and Ogives

3 Most Common Graphs in Research

1. **Histogram**
2. **Frequency Polygon**
3. **Cumulative Frequency Polygon (Ogive)**

Histograms, Frequency Polygons, and Ogives

The ***histogram*** is a graph that displays the data by using vertical bars of various heights to represent the frequencies of the classes.

The class boundaries are represented on the horizontal axis.

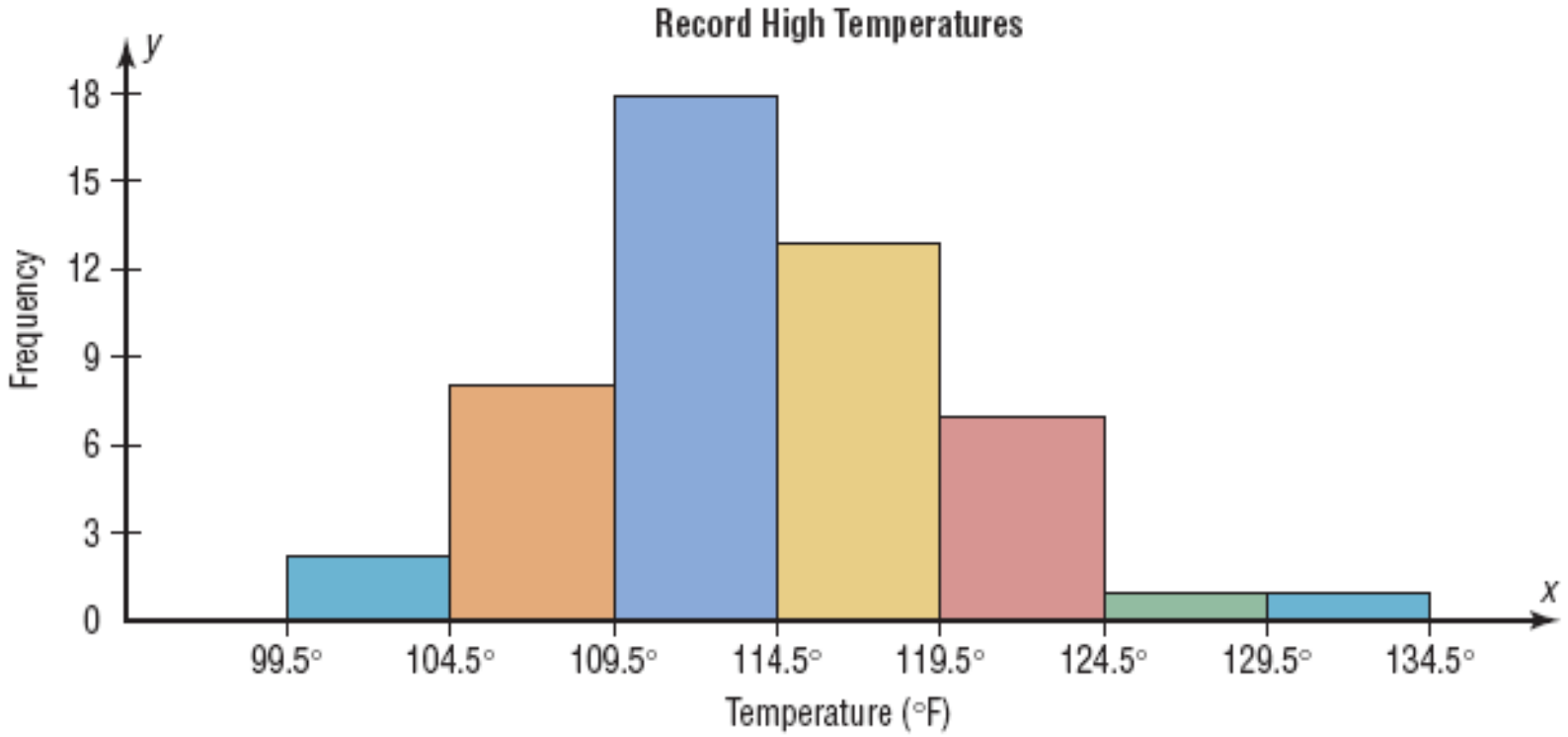
Histograms

Histograms use class boundaries and frequencies of the classes.

Class Limits	Class Boundaries	Frequency
100 - 104	99.5 - 104.5	2
105 - 109	104.5 - 109.5	8
110 - 114	109.5 - 114.5	18
115 - 119	114.5 - 119.5	13
120 - 124	119.5 - 124.5	7
125 - 129	124.5 - 129.5	1
130 - 134	129.5 - 134.5	1

Histograms

Histograms use class boundaries and frequencies of the classes.



2.2 Histograms, Frequency Polygons, and Ogives

- The ***frequency polygon*** is a graph that displays the data by using lines that connect points plotted for the frequencies at the class midpoints. The frequencies are represented by the heights of the points.
- The class midpoints are represented on the horizontal axis.

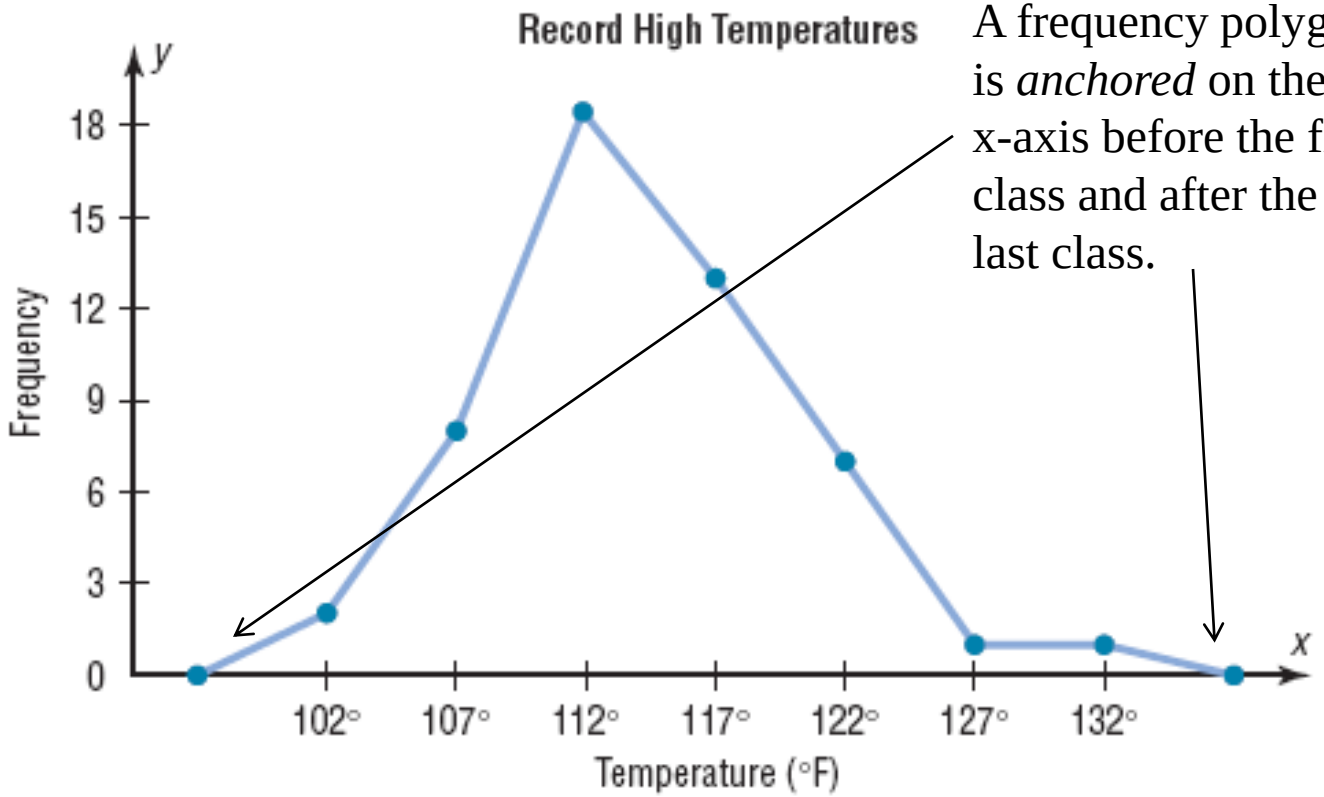
Frequency Polygons

Frequency polygons use class midpoints and frequencies of the classes.

Class Limits	Class Midpoints	Frequency
100 - 104	102	2
105 - 109	107	8
110 - 114	112	18
115 - 119	117	13
120 - 124	122	7
125 - 129	127	1
130 - 134	132	1

Frequency Polygons

Frequency polygons use class midpoints and frequencies of the classes.



A frequency polygon is *anchored* on the x-axis before the first class and after the last class.

2.2 Histograms, Frequency Polygons, and Ogives

- The ***ogive*** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

- The upper class boundaries are represented on the horizontal axis.

Ogives

Ogives use upper class boundaries and cumulative frequencies of the classes.

Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5	2	2
105 - 109	104.5 - 109.5	8	10
110 - 114	109.5 - 114.5	18	28
115 - 119	114.5 - 119.5	13	41
120 - 124	119.5 - 124.5	7	48
125 - 129	124.5 - 129.5	1	49
130 - 134	129.5 - 134.5	1	50

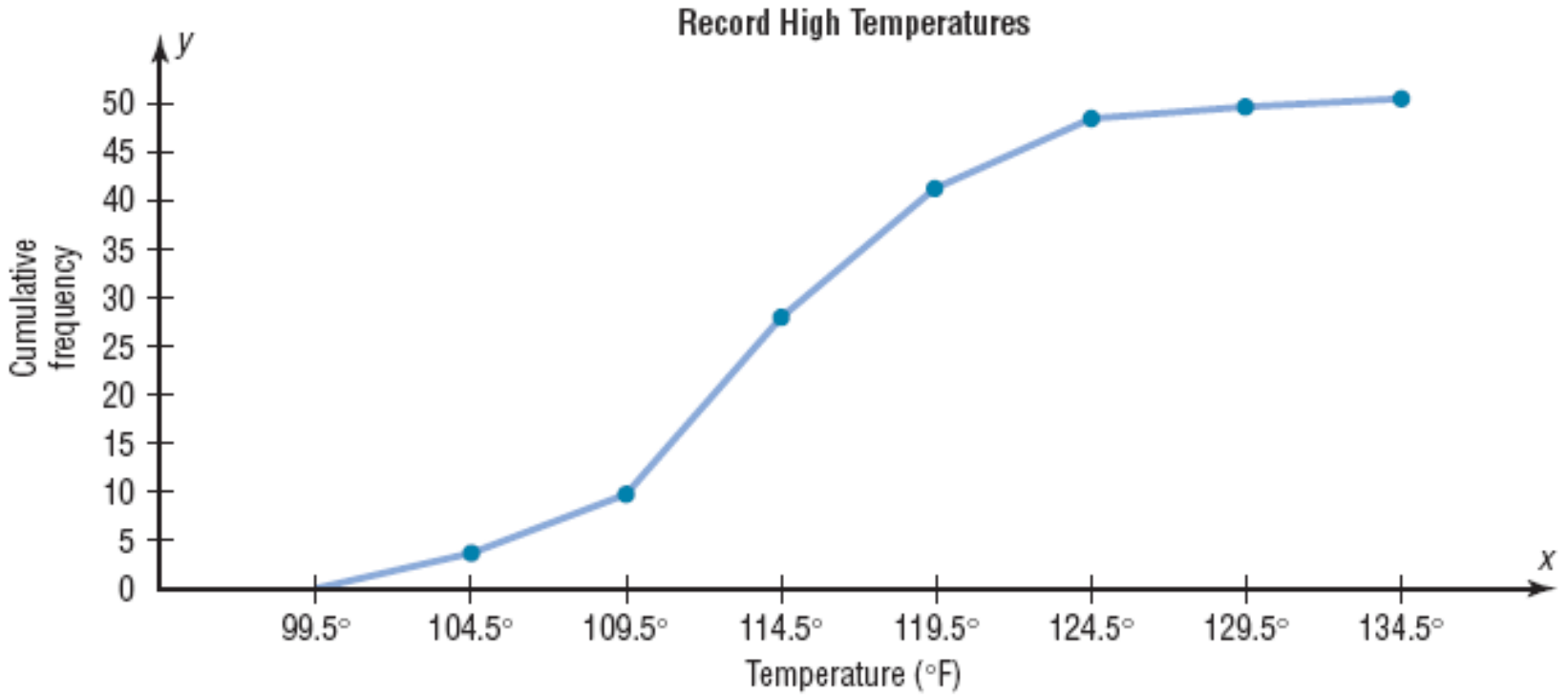
Ogives

Ogives use upper class boundaries and cumulative frequencies of the classes.

Class Boundaries	Cumulative Frequency
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Ogives

Ogives use upper class boundaries and cumulative frequencies of the classes.



Procedure Table

Constructing Statistical Graphs

1: Draw and label the *x* and *y* axes.

2: Choose a suitable scale for the frequencies or cumulative frequencies, and label it on the *y* axis.

3: Represent the class boundaries for the histogram or ogive, or the midpoint for the frequency polygon, on the *x* axis.

4: Plot the points and then draw the bars or lines.

Histograms, Frequency Polygons, and Ogives

If proportions are used instead of frequencies, the graphs are called ***relative frequency graphs***.

Relative frequency graphs are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class.

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week.

Class Boundaries	Frequency
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2

Histograms

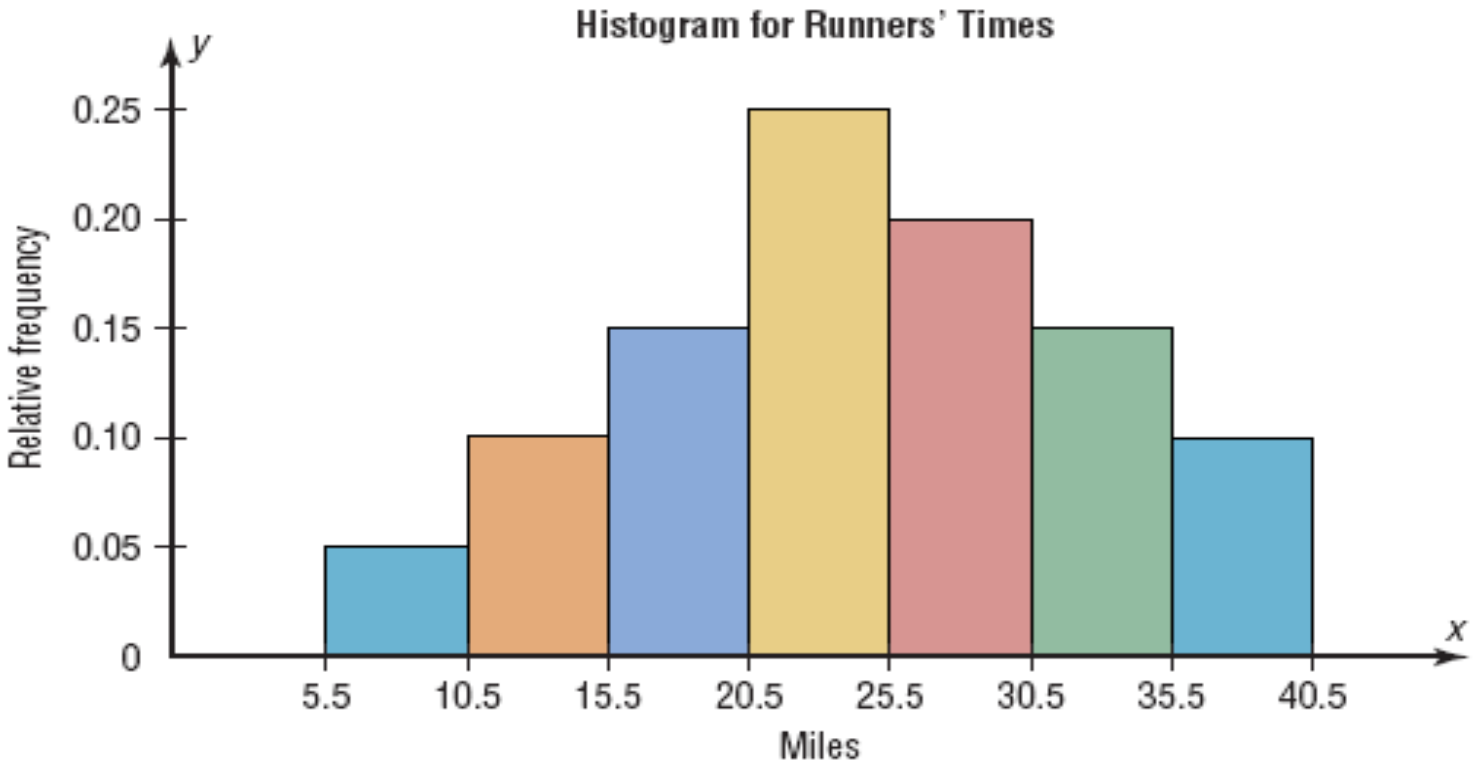
The following is a frequency distribution of miles run per week by 20 selected runners.

Class Boundaries	Frequency	Relative Frequency
5.5 - 10.5	1	1/20 = 0.05
10.5 - 15.5	2	2/20 = 0.10
15.5 - 20.5	3	3/20 = 0.15
20.5 - 25.5	5	5/20 = 0.25
25.5 - 30.5	4	4/20 = 0.20
30.5 - 35.5	3	3/20 = 0.15
35.5 - 40.5	2	2/20 = 0.10
	$\Sigma f = 20$	$\Sigma rf = 1.00$

Divide each frequency by the total frequency to get the relative frequency.

Histograms

Use the class boundaries and the relative frequencies of the classes.



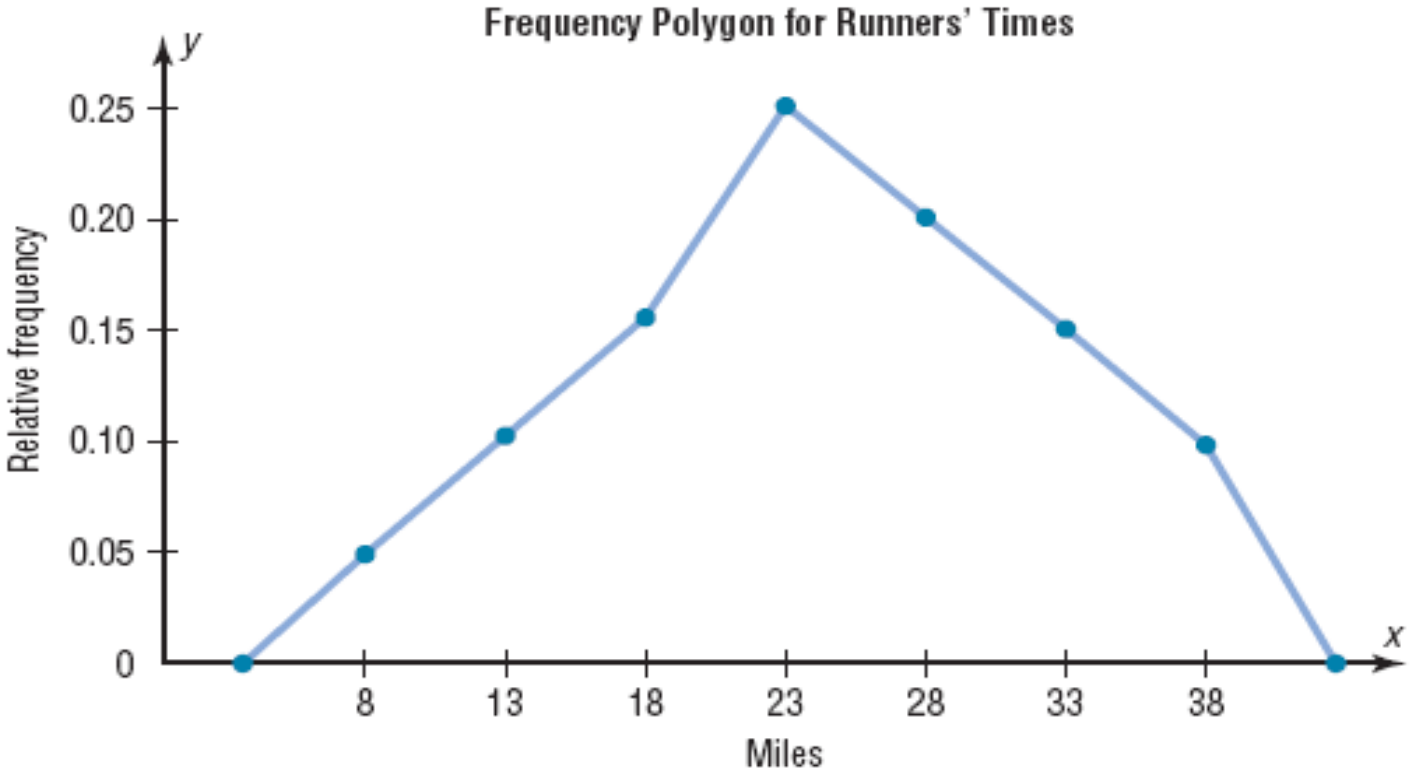
Frequency Polygons

The following is a frequency distribution of miles run per week by 20 selected runners.

Class Boundaries	Class Midpoints	Relative Frequency
5.5 - 10.5	8	0.05
10.5 - 15.5	13	0.10
15.5 - 20.5	18	0.15
20.5 - 25.5	23	0.25
25.5 - 30.5	28	0.20
30.5 - 35.5	33	0.15
35.5 - 40.5	38	0.10

Frequency Polygons

Use the class midpoints and the relative frequencies of the classes.



Ogives

The following is a frequency distribution of miles run per week by 20 selected runners.

Class Boundaries	Frequency	Cumulative Frequency	Cum. Rel. Frequency
5.5 - 10.5	1	1	1/20 = 0.05
10.5 - 15.5	2	3	3/20 = 0.15
15.5 - 20.5	3	6	6/20 = 0.30
20.5 - 25.5	5	11	11/20 = 0.55
25.5 - 30.5	4	15	15/20 = 0.75
30.5 - 35.5	3	18	18/20 = 0.90
35.5 - 40.5	2	20	20/20 = 1.00
	$\Sigma f = 20$		

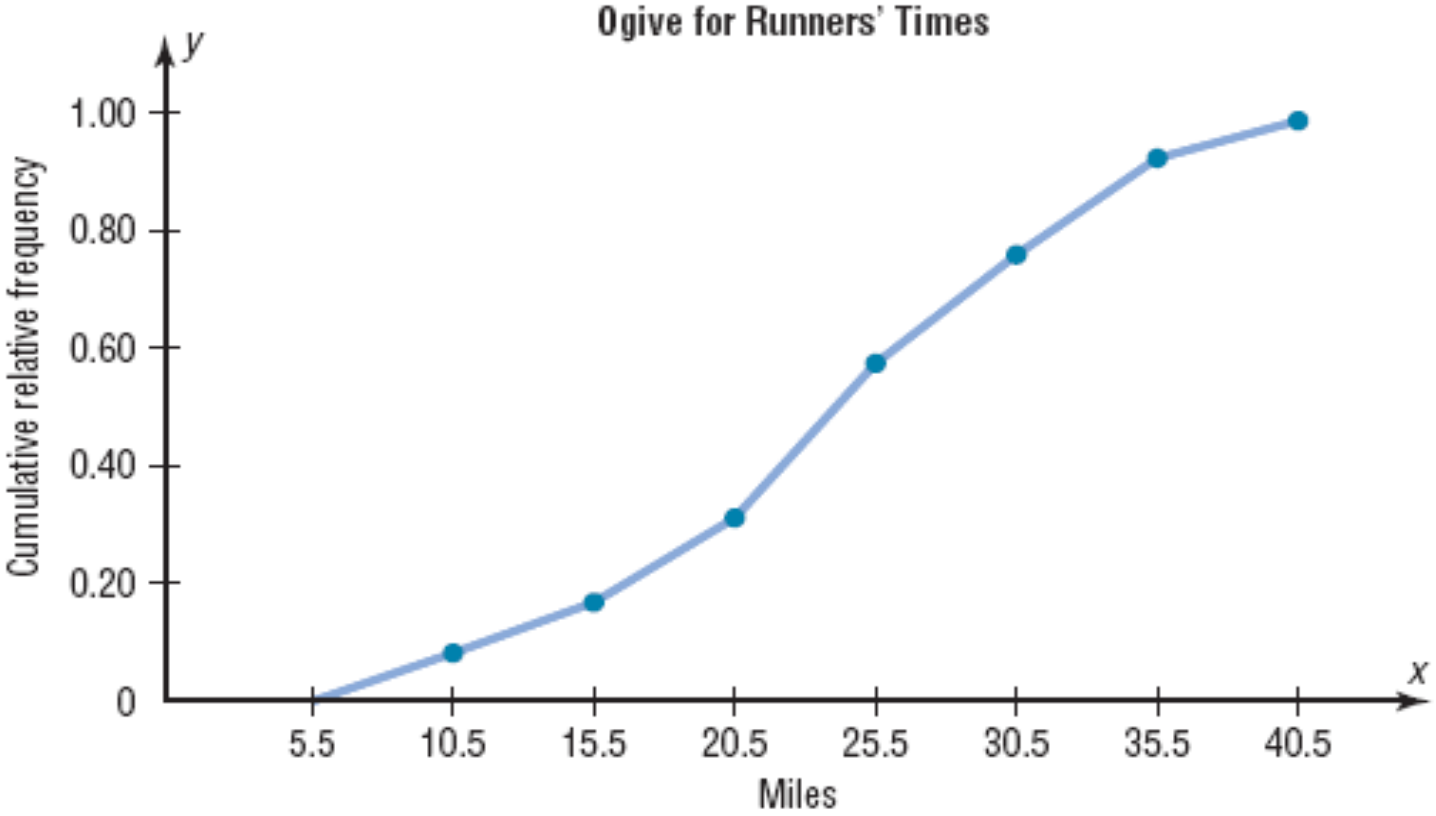
Ogives

Ogives use upper class boundaries and cumulative frequencies of the classes.

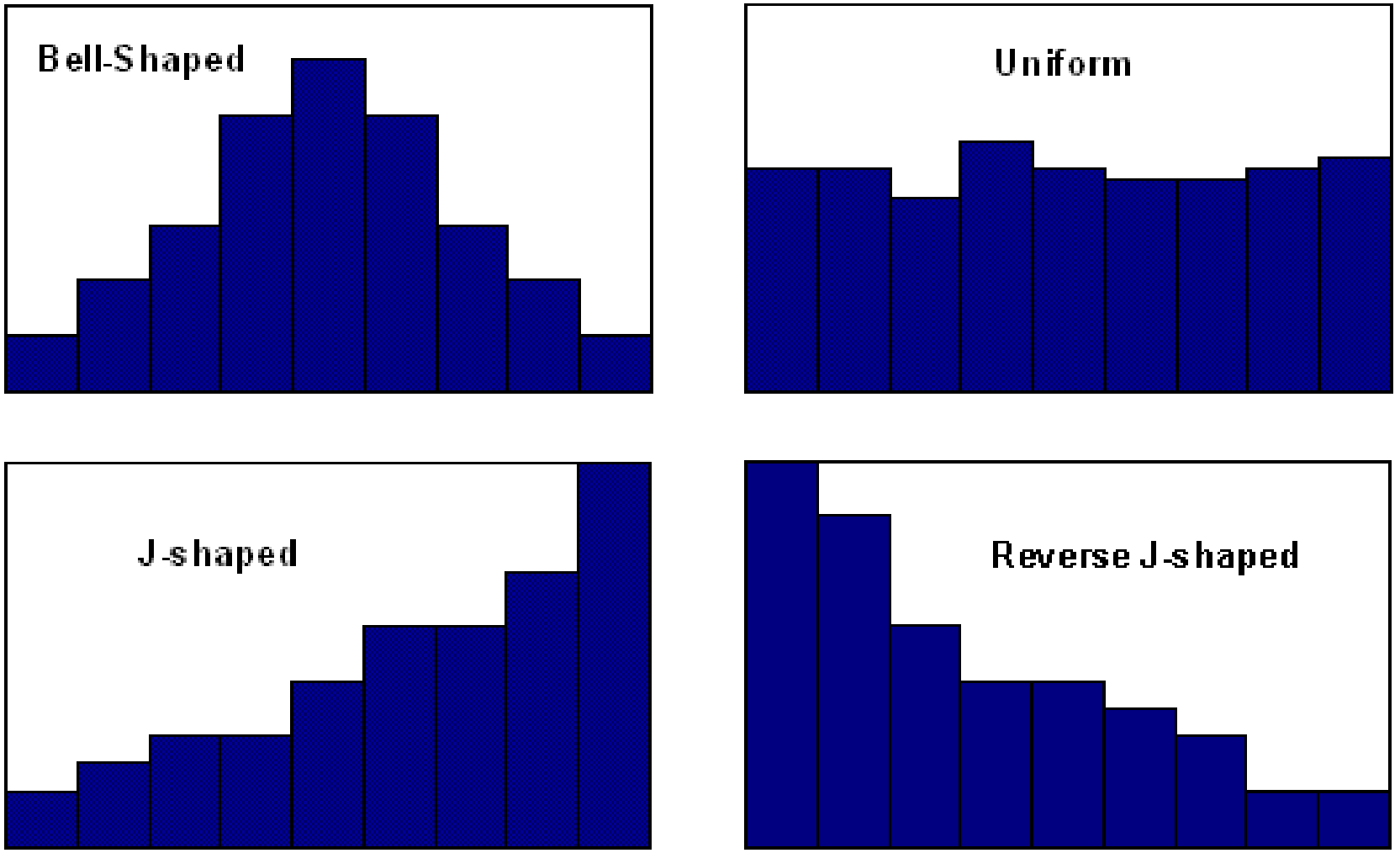
Class Boundaries	Cum. Rel. Frequency
Less than 10.5	0.05
Less than 15.5	0.15
Less than 20.5	0.30
Less than 25.5	0.55
Less than 30.5	0.75
Less than 35.5	0.90
Less than 40.5	1.00

Ogives

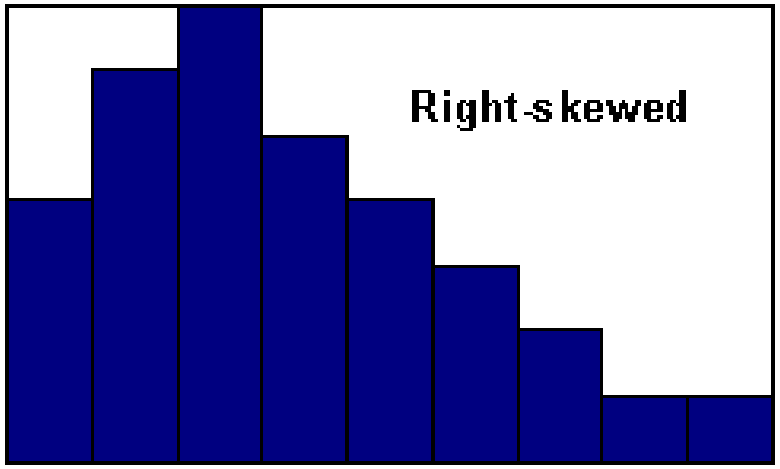
Use the upper class boundaries and the cumulative relative frequencies.

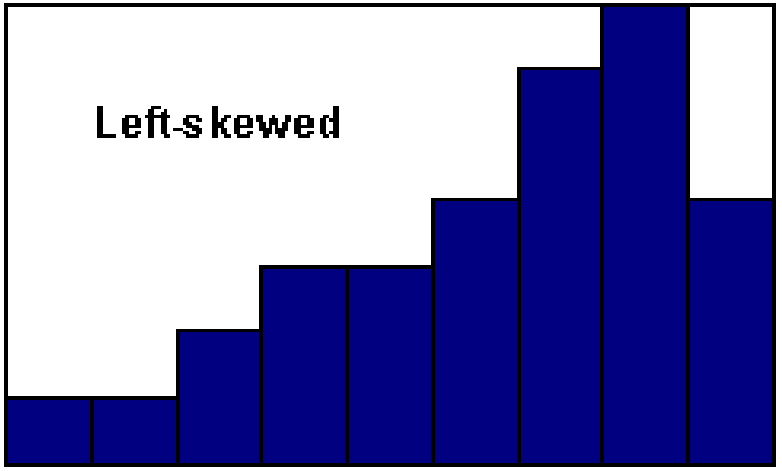


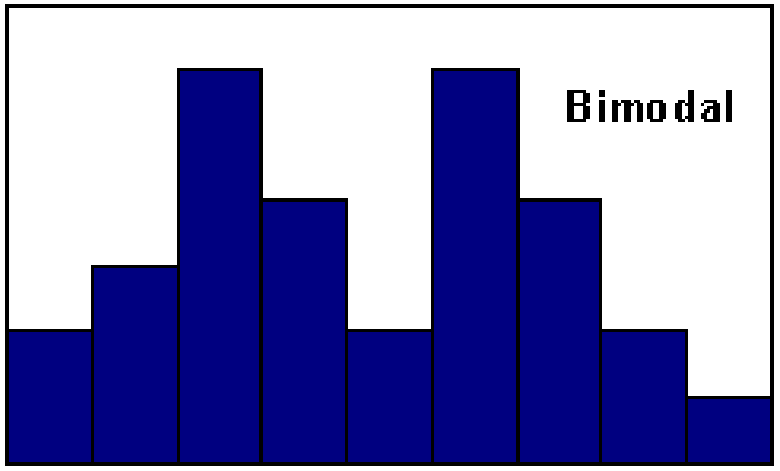
Shapes of Distributions

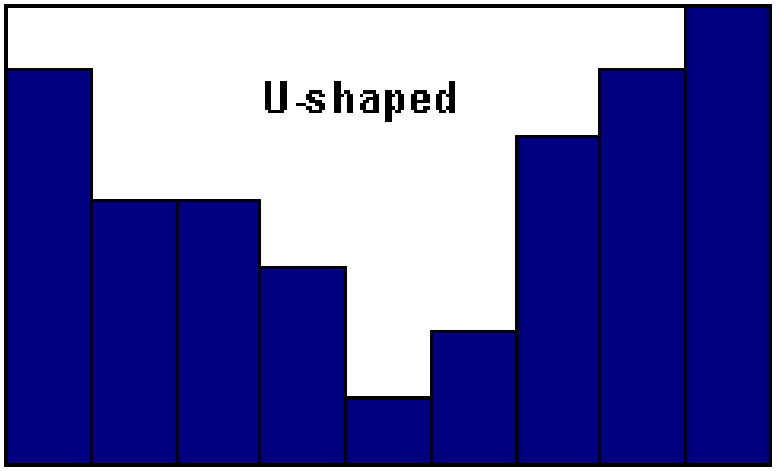


Shapes of Distributions









Thank You

CHAPTER 2

Frequency Distributions and Graphs

Dr. Hlaing Hlaing Moe

Professor/ Head

Department of Applied Statistics

Yangon University of Economics

Chapter 2 Overview

- 2-3 Other Types of Graphs

Chapter 2 Objectives

- Represent data using Pareto charts, time series graphs, and pie graphs.

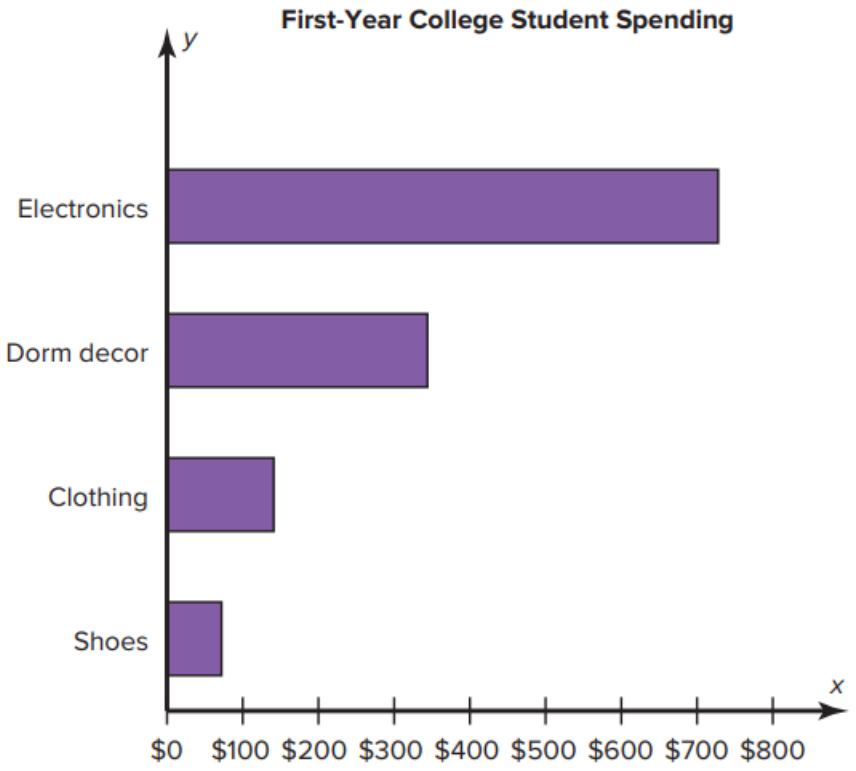
Bar Graphs

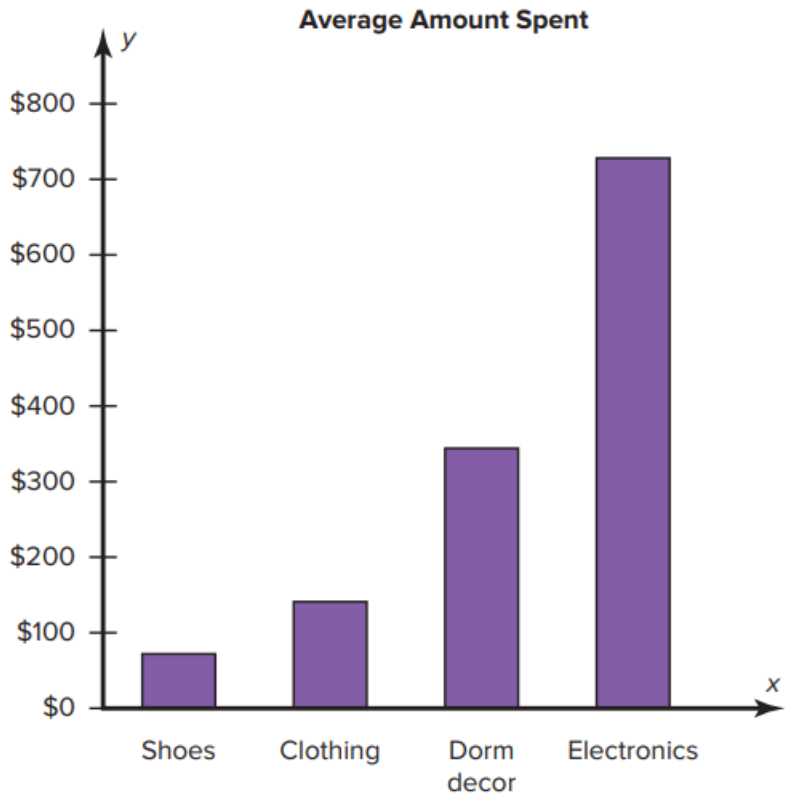
- When the data are qualitative or categorical, bar graphs can be used to represent the data.
- A bar graph can be drawn using either horizontal or vertical bars.
- A bar graph represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.
- Draw and label the x and y axes. For the horizontal bar graph place the frequency scale on the x axis, and for the vertical bar graph place the frequency scale on the y axis.

Horizontal Bar Graph

Vertical Bar Graph

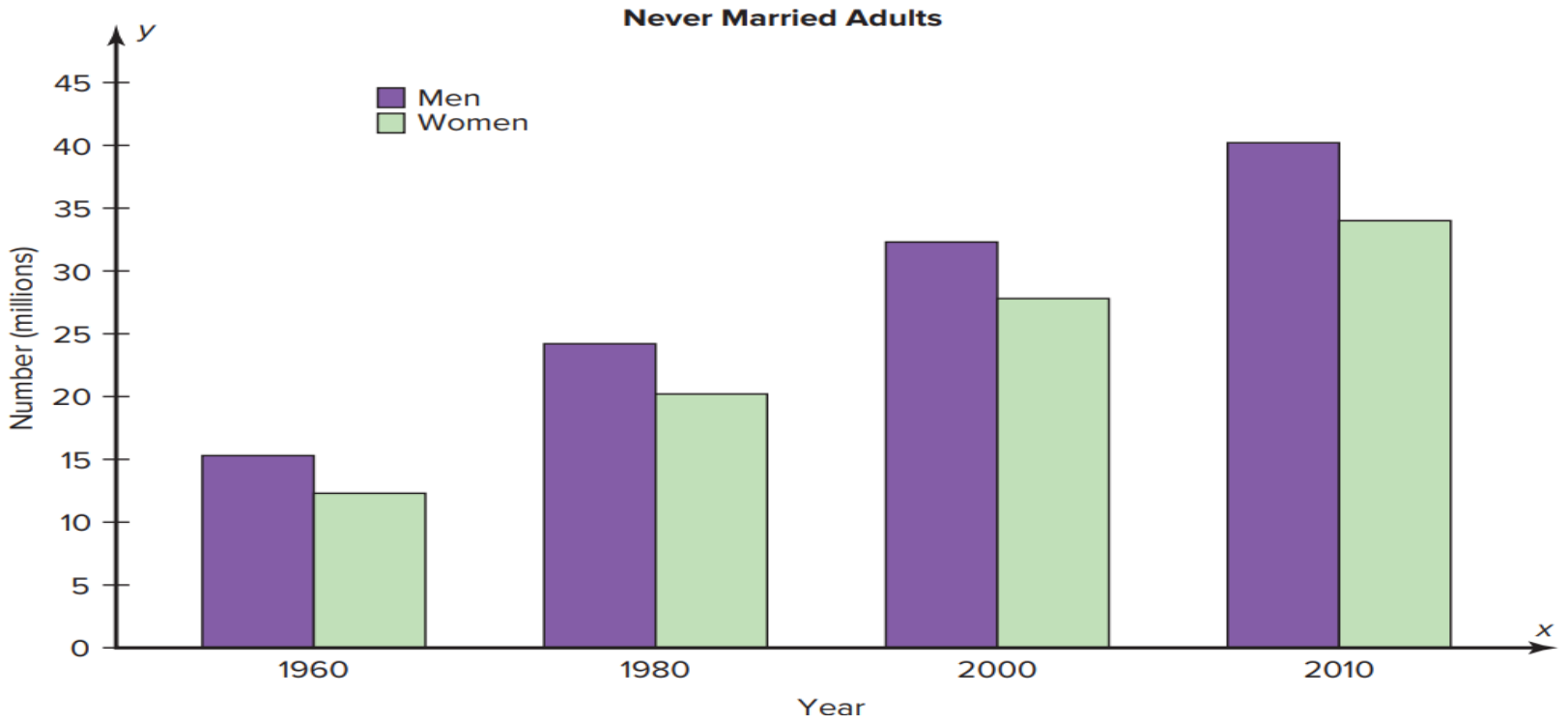
FIGURE 2–9 Bar Graphs for Example 2–8





Compound Bar Graphs

Bar graphs can also be used to compare data for two or more groups. These types of bar graphs are called compound bar graphs.



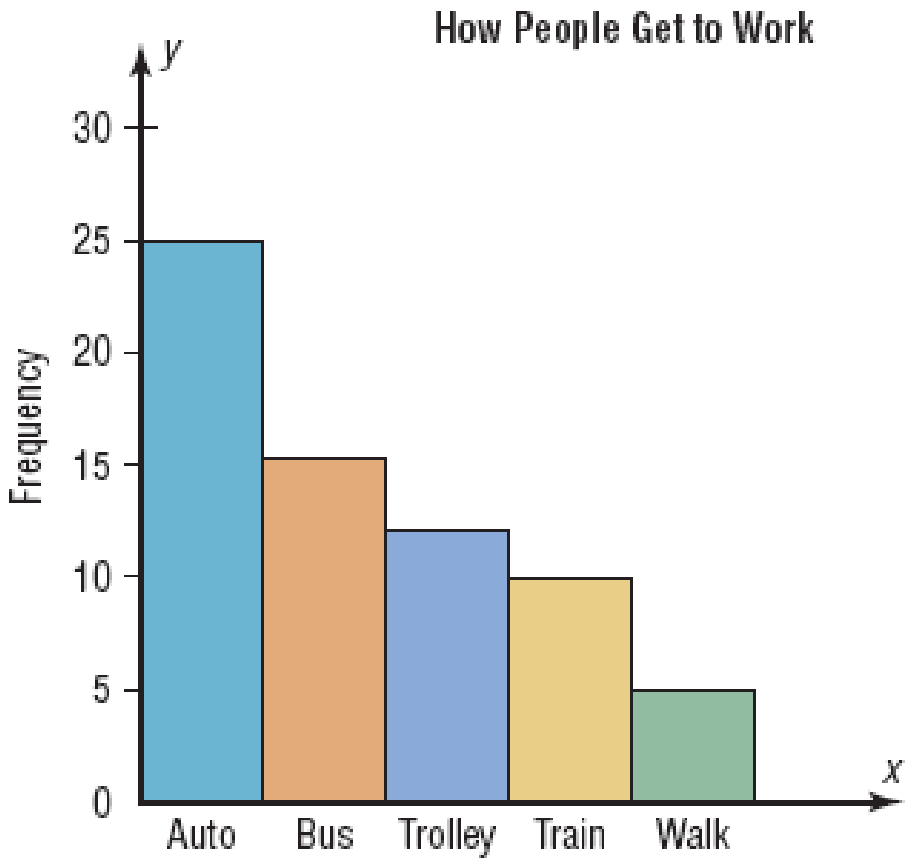
This graph shows that there have consistently been more never married males than never married females and that the difference in the two groups has increased slightly over the last 50 years.

Pareto Chart

- When the variable displayed on the horizontal axis is qualitative or categorical, a Pareto chart can also be used to represent the data.

- A Pareto chart is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

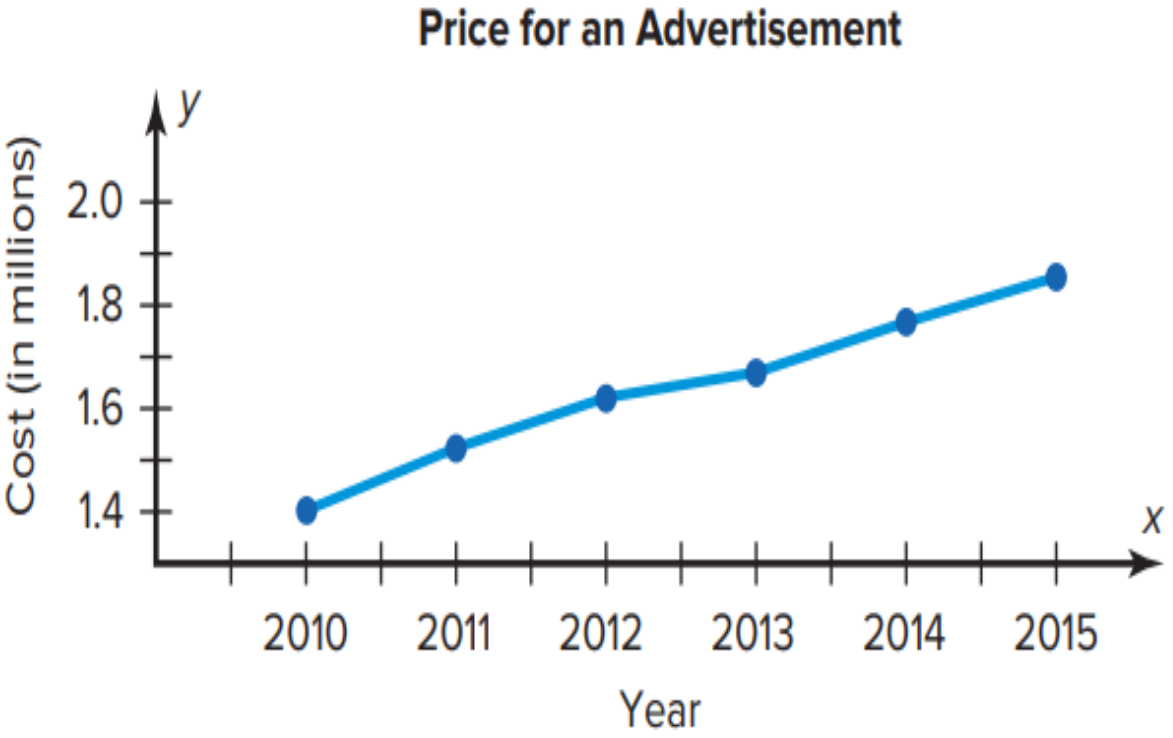
Pareto Charts



Time Series Graphs

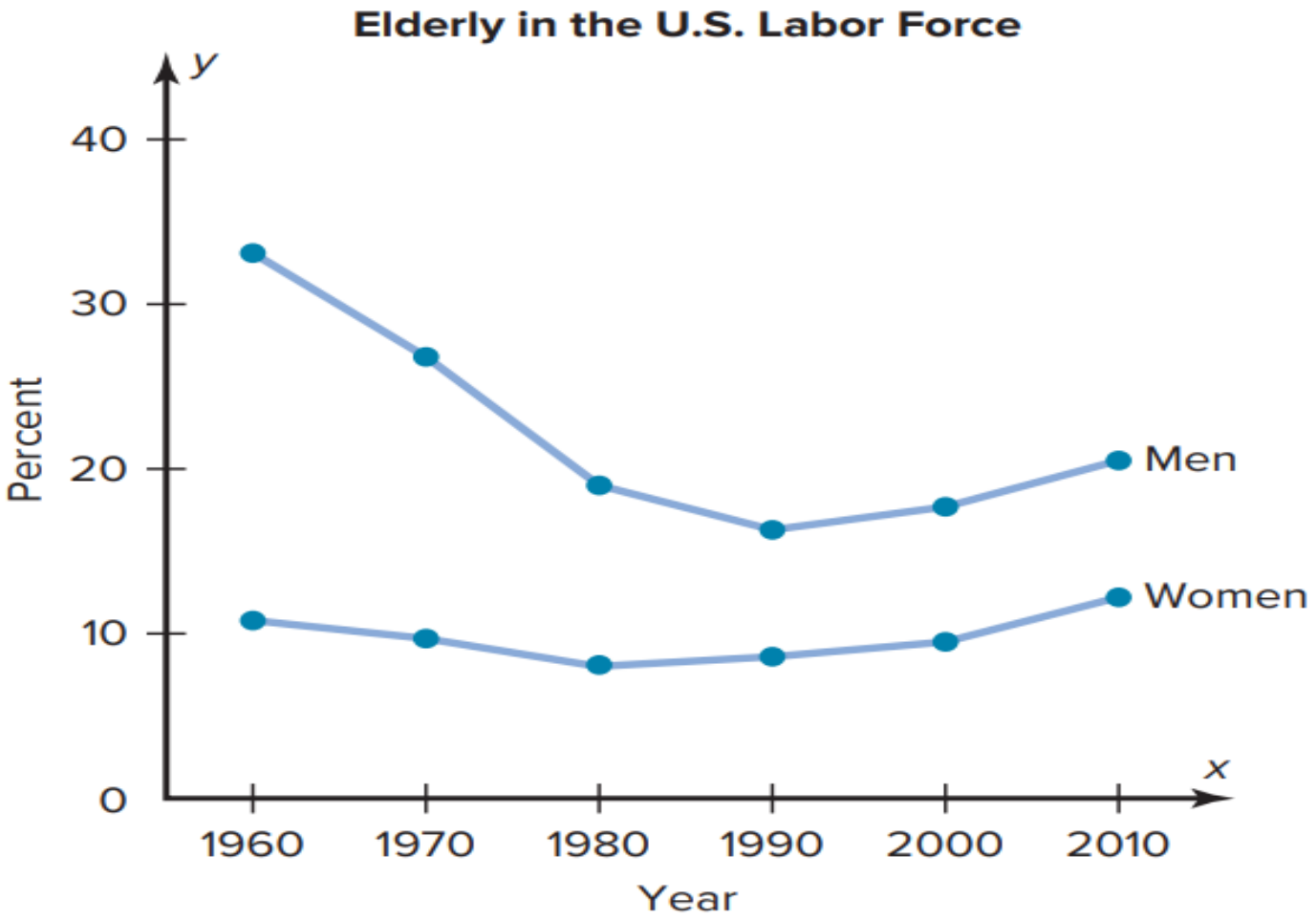
- When data are collected over a period of time, they can be represented by a time series graph.
- A time series graph represents data that occur over a specific period of time

FIGURE 2–12 Figure for Example 2–10



Compound Time Series Graph

Two or more data sets can be compared on the same graph called a compound time series graph if two or more lines are



Source: Bureau of Census, U.S. Department of Commerce.

Group Assignment I
Exercises 2.1

Island	No. 1 and 13
Ocean	No. 2 and 14
Beach	No. 3
Mountain	No. 4
Horizon	No. 5 and 15
Forest	No. 6 and 16
Sky	No. 17
River	No. 18

pgdrs.yue@gmail.com

Thank You

CHAPTER 2

Frequency Distributions and Graphs

Dr. Hlaing Hlaing Moe

Professor/ Head

Department of Applied Statistics

Yangon University of Economics

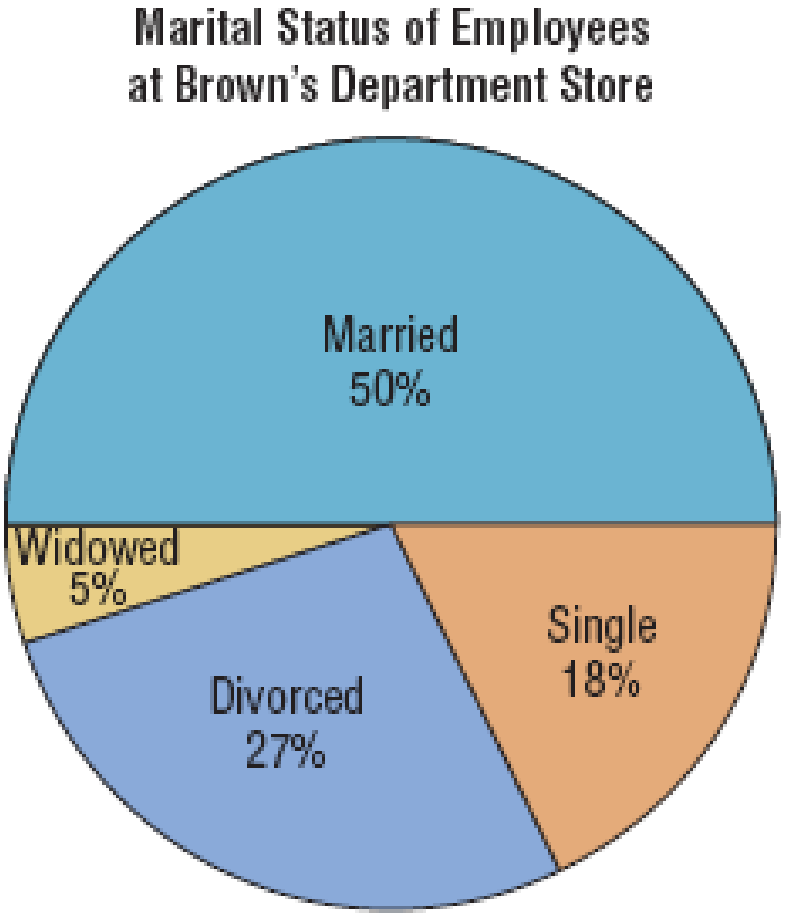
Chapter 2 Objectives

- Represent data using pie graphs, dotplots, and stem and leaf plots.

Pie Graphs

- Pie graphs are used extensively in statistics.
- The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections.
- Percentages or proportions can be used.
- The variable is nominal or categorical.
- A pie graph is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

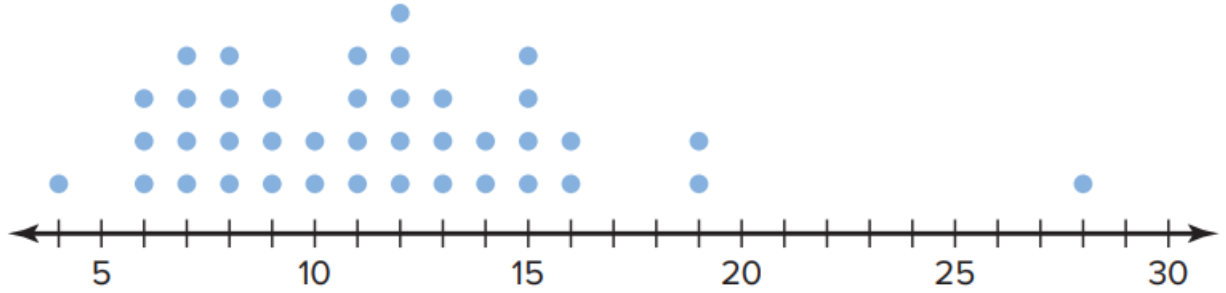
Pie Graphs


Marital Status of Employees
at Brown's Department Store
Married
50%
Widowed
5%
Single
18%
Divorced
27%

Dotplots

- A dotplot uses points or dots to represent the data values.
- If the data values occur more than once, the corresponding points are plotted above one another.
- A dotplot is a statistical graph in which each data value is plotted as a point (dot) above the horizontal axis.
- Dotplots are used to show how the data values are distributed and to see if there are any extremely high or low data values.

FIGURE 2–16 Figure for Example 2–13



The graph shows that the majority of the named storms occur with frequency between 6 and 16 per year. There are only 3 years when there were 19 or more named storms per year.

Stem and Leaf Plots

- The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing.
- It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.
- A stem and leaf plot is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	2	57	23
36	32	33	32	44
32	52	44	51	45

25	31	20	32	13
14	43	2	57	23
36	32	33	32	44
32	52	44	51	45

Unordered Stem Plot

0	2
1	3 4
2	5 0 3
3	1 2 6 2 3 2 2
4	3 4 4 5
5	7 2 1

Ordered Stem Plot

0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Excel Command

=REPT("0",COUNTIF($A$2:$A$11,F8*10+0))&REPT("1",COUNTIF($A$2:$A$11,F8*10+1))&REPT("2",COUNTIF($A$2:$A$11,F8*10+2))&REPT("3",COUNTIF($A$2:$A$11,F8*10+3))&REPT("4",COUNTIF($A$2:$A$11,F8*10+4))&REPT("5",COUNTIF($A$2:$A$11,F8*10+5))&REPT("6",COUNTIF($A$2:$A$11,F8*10+6))&REPT("7",COUNTIF($A$2:$A$11,F8*10+7))&REPT("8",COUNTIF($A$2:$A$11,F8*10+8))&REPT("9",COUNTIF($A$2:$A$11,F8*10+9))

Excel Command

=REPT(("9",COUNTIF($A$2:$A$11,F8*10+9))&REPT("8",COUNTIF($A$2:$A$11,F8*10+8
REPT("7",COUNTIF($A$2:$A$11,F8*10+7))&REPT("6",COUNTIF($A$2:$A$11,F8*10+6))&
EPT("5",COUNTIF($A$2:$A$11,F8*10+5))&REPT("4",COUNTIF($A$2:$A$11,F8*10+4))&R
PT("3",COUNTIF($A$2:$A$11,F8*10+3))&REPT("2",COUNTIF($A$2:$A$11,F8*10+2))&RE
T("1",COUNTIF($A$2:$A$11,F8*10+1))&REPT("0",COUNTIF($A$2:$A$11,F8*10+0))

Group Assignment III
Exercises 2.3

Island	No. 24, 15, 18
Ocean	No. 23, 22, 16
Beach	No. 11, 21, 13
Mountain	No. 10, 5, 14
Horizon	No. 4, 6, 17
Forest	No. 3, 7, 18
Sky	No. 2, 8, 19
River	No. 1, 27, 20

pgdrs.yue@gmail.com

Thank You

Yangon University of Economics

Department of Statistics

PGDRS-111

DATA STRUCTURE AND DATA COLLECTION METHODS

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Chapter 6

Methods of Data Collection

After reading this chapter, student should be able to:

- Understand the difference between qualitative and quantitative methods of data collection
- Describe various types of data collection methods, and state their uses and limitations
- Use an appropriate method or a combination of different methods for data collection
- Identify ethical issues involved in business research and the ways of ensuring that research informants or subjects are not harmed by the study

Chapter 6

Overview

6.1 Data Collection Method: Qualitative versus Quantitative

Quantitative Techniques

- 6.1.1 Observation
- 6.1.2 Survey Method
  - Personal Interview (Face to Face)
  - Mail Survey
  - Telephone Survey
  - Internet Survey

6 Introduction

- The data collection methods are the means to collect information about the objects under study in a systematic way.

- If data is collected randomly, it would be difficult to answer research questions in a conclusive way.
- Research can be divided into primary and secondary, based on the sources of data collection.
- Secondary research involves any information from published sources which has not been specifically collected for the current research problems. This includes the published sources of data such as an electronic database, periodicals, company's annual reports, etc.
- Primary data collection becomes necessary when a researcher is unable to find the data needed in secondary sources.
- Primary research involves collecting information specifically for the study in hand from the actual sources such as consumers, users/ non-users or other entities involved in the research.
- The primary data is therefore collected fresh and for the first time, and thus happens to be original in character.

6.1 Data Collection Method: Qualitative versus Quantitative

- Qualitative research explores attitudes, behavior and experiences through methods such as interviews or focus groups.
- It attempts to get an in-depth opinion from participants.
- As qualitative research examine attitudes, behavior and experiences which are related to the personal information of participants, fewer people take part in the research but the contact with these people tends to last a lot longer.
- There are many different methods such as participants observations, in-depth interview and focus group discussion.

6.1 Data Collection Method: Qualitative versus Quantitative

- Quantitative research generates statistics through the use of large-scale survey research, using methods such as questionnaires or structured interviews.
- If a business analyst or market researcher has stopped you on the streets, or you have filled in a questionnaire which has arrived through the post.
- This type of research reaches many more people, but the contact with those people is much shorter than it is in qualitative research.

6.1.1 Observation

- The observation method is the most commonly used method especially in studies relating to behavioural sciences.

- Observation is a technique that involves systematically selecting, watching and recording behavior and characteristics of living beings, objects or phenomena.

- Under the observation method, the information is sought by way of investigator's own direct observations without the knowledge of the respondents.

- The main advantage of this method is that the subjective bias is eliminated as actual behaviours get recorded and thus, we get more accurate information.

- Secondly, the information obtained under this method relates to what is currently happening; it is not complicated by either the past behavior or future intentions or attitudes.

- Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents.

While choosing observation as the method of data collection, the researcher should keep certain things in mind:

- What should be observed? How the observations should be recorded? Or how the accuracy of observations should be ensured?
- If the observation is characterized information of units to be observed, the style of recording the observed information, standardized conditions of observation and the selection of pertinent data of observation, then the observation is known as structured observation.
- On the other hand, if the observation is to take place without these characteristics to be sought in advance, it is known as unstructured observation.
- Structured observation is considered as appropriate in descriptive studies, whereas in an exploratory study, the observation is more likely to be unstructured.

However, the observational method has certain limitations.

- One of the major problem might be that we are not sure if a representative sample is chosen for the study, because recording of observational data take place at public places and we do not have control over who and how many are being observed at a given time.

6.1.2 Survey Method

- A survey involves the collection of information from representative target respondents using a predesigned questionnaire.
- Unlike observation, it is structured method of data collection in which we extract exactly the same information from all the target population.
- There are basically four types of survey used by researchers which are described below.

Personal interview (face to face)

- Among all the survey methods, personal interview (face-to-face) is the most widely used by researchers all over the world.
- These types of interviews consist of administering structured questionnaires and trained interviewers ask fixed, choice questions in a consistent format.
- In order to obtain a more accurate outcome of the overall survey, the interviewer should keep the following points in his/her mind. These important tips will ensure reliable, credible and unbiased responses from the respondents.

Mail survey

- Imagine that you are interested in exploring the attitudes college students have about writing.

- Since it would be impossible to interview every student on campus, choosing the mail-out survey as your method would enable you to choose a large sample of college students.

- You might choose to limit your research to your own college or university, or you might extend your survey to several different institutions.

- If your research question demands it, the mail survey allows you to sample a very broad group of subjects at a small cost.

Telephonic survey

- This is an alternative form of interview to the personal, face-to-face interview, where the interviewer collects the relevant information from the target respondents through telephonic conversation.

- Following points should be helpful to locate the respondents and make them agree to participate for the telephonic survey.

Internet survey

- With the growth of the Internet and the expanded use of electronic mail for all purposes, the electronic survey is becoming one of the most widely used survey method these days.

The electronic surveys can be done in many ways:

- (i) The survey forms can be distributed as electronic mail messages through attachment to potential respondents,
- (ii) the survey form can be posted as World Wide Web forms on the Internet, and
- (iii) the survey form can be distributed via publicly available computers in high-traffic areas such as libraries and shopping malls.

Thank You

Yangon University of Economics

PGDRS-111

Data Structure and Data Collection Methods

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Chapter 6

Methods of Data Collection

After reading this chapter, student should be able to:

- Understand the difference between qualitative and quantitative methods of data collection
- Describe various types of data collection methods, and state their uses and limitations
- Use an appropriate method or a combination of different methods for data collection
- Identify ethical issues involved in business research and the ways of ensuring that research informants or subjects are not harmed by the study

Chapter 6

Overview

- 6.1.3 Qualitative Techniques
  - In-depth interview
  - Focus group discussion
  - Projective techniques
- 6.2 Collection of Secondary Data
- 6.3 Selection of Appropriate Methods of Data Collection
- Ethical Issues

6.1.3 Qualitative Techniques

- The research objective calls for more indirect methods of questioning, either because normal quantitative surveys are inadequate or inappropriate.
- In such cases, qualitative methods, which probe the mind of respondent, might be useful.
- The major requirement for qualitative techniques is that we need a behavioural specialist such as a psychologist or sociologist to analyze the findings.
- The sample size in qualitative techniques is usually small, and analysis and interpretation is not as easy as it is in the quantitative studies.
- If done by non-expert, qualitative research can be completely misleading. Some of the important tools are discussed in detail below.

In-depth interview

- In-depth interviewing is a qualitative research technique that involves conducting intensive individual with a small number of respondents to explore their perspectives on a particular idea, programmer or situation.
- In- depth interviews are useful when detailed information about a person's thoughts and behaviours is needed or new issues are to be explored in depth.
- Interviews are often used to provide context to other data (such as outcome data), offering a complete picture of what happened in the programme and why.
- In-depth interviews should be used in place of focus groups if the potential participants are not to be included or are not comfortable talking openly in a group, or when one wants to distinguish an individual's (as opposed to group) opinions about the programme.
- They are often used to refine questions for future surveys of a particular group.
- The primary advantage of in-depth interviews is that they provide much more detailed information than what is available through other data collection methods such as surveys.

Focus group discussion

- A focus group is a group of interacting individuals having some common interests or characteristics, brought together by a researcher, who uses the group and its interaction as a way to gain information about a specific or focused issue.
- The focus group is a carefully planned and moderated discussion to obtain the meaningful information on the area of interest in a non-threatening environment.
- The focus group discussion is an unstructured method of data collection where the respondents express their views freely.
- It is mostly used for explorative studies rather than any conclusive studies.
- Groups are comprised for respondents who share similar concerns and responsibilities but have minimal contact with each other in their daily lives.
- As groups differ in their composition and dynamics, multiple groups should be organized to obtain information from a different perspective on a given topic.
- Groups typically contain approximately 6-12 people: large enough to provide for a range of views but small enough for everyone to contribute.

Projective techniques

- Projective techniques are used by psychologists of respondents in inferring underlying motives, urges or intentions which are such that the respondents resist revealing them or are unable to figure it out themselves.

- The respondent in supplying information tends unconsciously to project his/her own attitudes or feelings on the subject under study.

- Projective techniques play an important role in motivational studies or in attitude surveys. In the following paragraphs, we shall discuss some of the important projective techniques.

6.2 Collection of Secondary Data

- Secondary data is indispensable for most organizational research.
- Secondary data refers to information that have been already gathered by someone (individual or agencies) and readily available to the researcher.
- The secondary data can be internal or external to the organization and it can be accessed through the Internet or perusal of recorded or published information.
- Generally, business research should be undertaken after a prior search of secondary sources.
- The secondary sources of information are important for any business research due to the following reasons.
  - Secondary data may be available which is entirely appropriate and wholly adequate to draw conclusions and answer the question or solve the problem. Sometimes primary data collection simply is not necessary.
  - It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget, a thorough examination of secondary sources can yield a great deal of more information than can be gathered through a primary data collection exercise.

6.2 Collection of Secondary Data

- The time involved in searching secondary sources is much less than that needed to complete primary data collection exercise.
- Secondary sources of information can yield more accurate data than that obtained through primary research.
- It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses.
- Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken.

6.2 Collection of Secondary Data

- There are several sources of secondary published data and available in
  - (a) various publications of the central, state and local governments such as publications of economic indicators, census data and statistical abstracts;
  - (b) various publications of foreign governments or international bodies and their subsidiary organizations;
  - (c) books and periodicals;
  - (d) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc;
  - (e) reports generated by research scholars of academic institutions in various fields; and
  - (f) public records and statistics, historical documents and other sources of published information.
  - The secondary data also could be from unpublished sources such as unpublished biographies and autographies, unpublished research thesis, working research papers of scholars, etc.

6.2 Collection of Secondary Data

- Researchers must be very careful in using secondary data because it is just possible that the available data may not be suitable or may be inadequate in the context of the problem under investigation. By way of caution, the researcher, before using secondary data, must see that they possess following characteristics.
- Reliability of data: The reliability can be tested by investigating such things about the said data: (a) Who collected the data? (b) What were the sources of data?(c) Was the data collected by using appropriate methods? (d) At what time period data was collected? (e) What level of accuracy was desired and how far it was achieved?
- Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable for another enquiry. The researcher must carefully scrutinize the definition of various terms and units of collection used in the study before identification of relevant data from the published sources.
- Adequacy of data: If the level of accuracy in data is found inadequate for the purpose of present enquiry, data will be considered as inadequate and should not be used by a researcher. The data will also be considered as inadequate, if they are related to an area which may be either narrower or wider than the area of the present study.

6.3 Selection of Appropriate Methods of Data Collection

- We have discussed various methods of data collection. Each method has its own advantages and disadvantages. The researcher must judiciously select the method(s) for his/her own study keeping in view the following factors.
  - Nature, scope and object of enquiry
  - Availability of funds
  - Time factor
  - Precision required

6.4 Ethical Issues

- Several ethical issues should be addressed while collecting data. We should be concern on whether one's procedures of collecting information are likely to cause any physical or emotional harm. These may be caused by:
- violating participants' right to privacy by posing sensitive questions or by gaining access to records which may contain personal data;
- observing the behavior of participants without their being aware
- allowing personal information to be made public which participants would want to be kept private; and
- failing to observe/ respect certain cultural values, traditions or taboos valued by the participants.

6.4 Ethical Issues

- Several methods for dealing with these issues may be recommended:
- obtaining the respondent's consent before the study or the interview begins;
- not exploring sensitive issues before a good relationship has been established with the participant;
- ensuring the confidentiality of the data obtained; and
- learning enough about the culture of participants to ensure it is respected during the data collection process.
- If sensitive questions are asked, for example, about family planning or sexual practices, or about opinions of patients on the health services provided, it may be advisable to omit names and addresses from the questionnaires.

Thank You

Measurement in Management and Scaling Techniques

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Learning Outcomes

After reading this chapter, student should be able to:

- Understand the four basic measurement techniques in business research
- Learn different measurement scales under comparative and non-comparative scaling techniques. Select the correct measurement scales for different types of statements in the questionnaire and take a number of practical decisions into account while developing the scale for questions
- Test the measurement instruments for its degree of stability, consistency and reliability

Introduction

- Measurement is fundamental to business research. Without appropriate measurement, it is difficult, if not impossible, to comment on business behavior or business phenomenon.
- A scale of measurement allows the investigator to make comparisons of amounts and changes in the variable being measured.
- Measurement consists of two basic processes called conceptualization and operationalization.

Introduction (Continued)

- First, variables are defined by conceptual definitions (constructs) that explain the concept the variable is attempting to capture.

- Secondly, variables are defined by operational definitions, that is, definitions of how variables will be measured.

- It is followed by an advanced process called determining the levels of measurement, and measuring reliability and validity of an instrument which is the focus of the current chapter.

Basic Measurement Techniques

- Level of measurement is important while measuring the variables.
- The higher the level of measurement of a variable, the more powerful is the statistical techniques that can be used to analyze it.
- There are four basic types of measurement in business research. They are nominal, ordinal, interval and ratio.

Nominal Scale

Nominal measurement consists of assigning items to groups or categories.

It has following features:

- It is used to indicate categories.
- Numbers are only used as labels.
- It has no numerical significance.
- It does not represent any order or distance.

- ✓ Nominally scaled variables cannot be used to perform many statistical computations such as mean and standard deviation, because such statistics do not have any meaning when used with nominal scale variables.
- ✓ With nominally scaled variables, the analysis is confined to frequency and cross-tabulation.
- ✓ The chi-square test can be performed on a cross-tabulation of nominal scale data.

Ordinal Scale

- The main characteristic of the ordinal scale is that the categories have a logical or ordered relationship to each other.
- These types of scale permit the measurement of degrees of difference, but not the specific amount of difference.
- This scale is very common in marketing, satisfaction and attitudinal research.

i. Ordinal scale implies "ranking" on the basis of preference.
ii. It does not say anything about the "distance".
iii. Ranks are not interchangeable as nominal scale labels are.

Ordinal Scale (Continued)

- In addition to frequency tabulation and cross- tabulations, the other statistics that can be used with the ordinal scale are median, various percentiles such as quartile and rank correlation.
- The arithmetic mean should not be used as the average ranking does not make any sense here.

Interval scale

- The interval scale is also known as rating scale and variables (attributes) are measured on different scales such as scale of 1 to 5 or 1 to 7 or 1 to 10.
- In this scale, it is assumed to have equidistant points between each of the scale elements. This means that we can interpret differences in the distance along the scale.
- The interval data can be used to calculate mean, standard deviation, correlation coefficient, regression, analysis of variance, factor analysis, and a whole range of advanced multivariate and modeling techniques.

Ratio Scale

- A ratio scale possesses all the properties of the nominal, ordinal and interval scale, and in addition, an absolute meaningful zero point. (Example- age)
- It is the top level of measurement and it can be used for all the statistics as in the case of interval scale.


Thank you
for your
attention

Measurement in Management and Scaling Techniques

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Advanced Measurement Techniques

- The various types of scaling techniques used in research can be classified into two categories: (i) comparative scales and (ii) non-comparative scales.
- In comparative rating scale, the respondent evaluates (assign a value) to a particular item/brand/product in comparison to other items/brands/products.
- The evaluation cannot be possible without comparison.
- On the other hand, in non-comparative scaling respondents evaluate each item/brand/product independently without comparing with any other items/brands/products.

Comparative Scales

The comparative scales can be further divided into the following four types of scaling techniques:

(i) paired comparison scale

(ii) rank order scale

(iii) constant sum scale and

(iv) Q-sort scale.

Paired Comparison Scale

- This is a comparative scaling technique in which a respondent is presented with two objects at a time and asked to select one object (rate between two objects at a time) based on some criterion.
- The data obtained is ordinal in nature.
- A respondent has to make [n (n-1)/2] paired comparison for n number of items.

Advantages

- Some special techniques such as multidimensional scaling require the data to be collected based on pair comparison.

Disadvantages

- Data obtained is ordinal.
- If the number of objects (n) is large, there is a high risk of ill considered answers or refusal or refusal to answer.

Rand Order

A respondent is presented with several objects simultaneously and ask to order or rank them according to some criterion.

Advantages

- It takes less time as compared to "paired comparison scaling". If there are "n" stimulus objects, only (n-1) scaling decisions need to be made.
- Rank order scaling is commonly used to measure preference of the brand as well s attributes.
- Rank order data is obtained in ordinal data.

Disadvantages

- Rank order scaling results in ordinal data.

Constant Sum Scaling

- A respondent allocates a constant sum of units (usually points) among a set of stimulus objects with respect to some criterion.
- If an attribute is unimportant, the respondent may assign it zero point.
- If an attribute is twice as important as some other attribute, it receives twice as many points.

Advantages

- It allows fine discrimination among stimulus objects without requiring too much time.
- The constant sum scaling not only measures the preference but also the degree to which a particular attribute is more important than others.

Disadvantages

- Respondents may allocate more or fewer units than those specified.

Q-Sort Scale

- It uses a rank order procedure and the objects are sorted into piles based on similarity with respect to some criteria.
- The number of objects to be sorted varies between 60 and 140 approximately.
- Let us say there are nine brands. On the basis of taste we can classify the brands into tasty, moderate and non-tasty. We can also classify on the basis of price as low, medium and high. Then we can attain the perception of people that whether they prefer low-priced brand, high or moderate. We can classify 60 brands or pile it into 3 piles. So the number of objects is to be placed in three piles-low, medium or high.
- Thus, the Q-sort technique is an attempt to classify subjects in terms of their similarity to attribute under study.

Non-Comparative Scales

Non-comparative scales are further divided into following four-types:

(i) Continuous rating scale

(ii) Likert scale

(iii) Semantic differential scale and

(iv) Stapel's scale

Continuous Rating Scale

- This is the oldest and most widely used method for performance appraisal.
- A respondent is asked to rate the objects by placing a mark at the appropriate position on a line that runs from one extreme criterion variable to other.
- A respondent is not restricted to selecting from marks previously set by the researcher. It is also known as graphic rating scale.
- Here, respondents do not necessarily need to choose the predefined point.
- They are free to occupy any position in the graph.

Itemized Rating Scale

Likert Scale

- It is one of the most popular non-comparative rating scaling techniques in management research.
- In this scale, the respondent indicates a degree of agreement or disagreement with each of the series of statements about the stimulus objects.
- Each statement is assigned a numerical score ranging from either-2 to +2 or 1 to 5. Total score of each respondent is calculated by summing across the item.

Advantages

- Likert scale is easy to construct and administer.
- Respondents readily understand how to use the scale, making it suitable for mail, telephone or personal interview.

Disadvantages

- It takes longer time to complete than other itemized rating scales because respondent has to read each statement.
- Care needs to be taken when using Likert scales in cross-cultural research, as there may be cultural variations in willingness to express disagreement.

Semantic Differential Scale

- Another scale that is commonly used by business researchers is the semantic differential scale.
- This is quite similar to rating scale in which, the end-points are associated with bipolar labels (adjectives) such as "cold" and "warm" or "unreliable" and "reliable", and so on.
- There are many intermediate points in between two extreme points and could be coded as 1 to 5 or 1 to 7.

Advantages

- The advantage of using semantic differential is its simplicity, while producing results comparable with those of the more complex scaling methods.

- The method is easy and fast to administer, but it is also sensitive to small differences in attitude, highly versatile, reliable and generally valid.

Stapel's Scale

Stapel's scale, developed by Jan Stapel, is useful for researchers to understand the positive and negative intensity of attributes of respondents.

It has following distinctive features:

- Each item has only one word/phrase indicating the dimension it represents.
- Each item has ten response categories.
- Each item has an even number of categories.
- The response categories have numerical labels but no verbal labels.

Practical Consideration

- While developing the scale for questions, a number of practical decisions should be taken into account.
- They are described below.

(i) Number of scale categories

(ii) Number of items to measure concept

(iii) Odd or even number of categories

(iv) Balanced scale or unbalanced scale

(v) Forced choice or non-forced choice

Number of Scale Categories

- One of the important decisions that one has to consider is whether one should have 5 point scale or 7 point scale or 10 point scale.
- From a research design perspective, the larger the number of categories, the greater the precision of the measurement scale.

Number of Items to Measure Concept

- Concepts are measured using scales with multiple items known as multi-item scales.
- A multi-item scale consists of a number of closely related individual statements (items) whose responses are combined into a composite score.
- Generally it is common to see five to seven items and even more to measure a single concept. A minimum of three items is must to achieve acceptable reliability.

Odd or Even Number of Categories

- There is debate on whether one should use even or odd numbers of points on a rating scale.
- In an odd number of categories of a scale, the mid-point represents a neutral position.
- This type of scale is generally used when based on the experience or judgement of the researchers, it is believed that a part of the sample are likely to feel neutral about the issue being examined.
- However, there are reasons a researcher might prefer an even scale over an odd one as given below.

Strongly disagree **Disagree** **Agree** **Strongly agree**

○ ○ ○ ○

Strongly disagree **Disagree** **Neither agree nor disagree** **Agree** **Strongly agree**

○ ○ ○ ○ ○

Balanced Scale or Unbalanced Scale

- Scales can be either balanced or unbalanced.
- The scale is known as balanced if the number of positive (favourable) options is equal to number of negative (unfavourable) options.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
○	○	○	○	○

- On the other hand, the scale is known as unbalanced if the number of positive options is greater than the number of negative potions or the number of positive options is less than the number of negative options.
- The scales given below are examples of unbalanced scale.

Strongly disagree	Disagree	Slightly disagree	Agree	Strongly agree
○	○	○	○	○

Strongly disagree	Disagre	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
○	○	○	○	○	○

Forced Choice or Non-forced Choice

- The discussion on even-point scales leads us into a discussion of forced and unforced choice questions.
- The four-point scale below is an example of forced choice.

Strongly disagree	Disagree	Agree	Strongly agree
○	○	○	○

- In this case, the respondents are forced to either agree or disagree.
- However, the five-point scale below is an example of unforced choice question.

Strongly disagree	Disagree	Neither agree disagree	Agree	Strongly agree
○	○	○	○	○

The Characteristics of Good Measurement

- Sound measurement must meet the tests of validity, reliability and practicality.
- In fact, these are the three major considerations one should use in evaluating a measurement tool.
- Validity refers to the extent to which a test measures what we actually wish to measure.
- Reliability has to do with the accuracy and precision of a measurement procedure.
- Practicality is concerned with a wide range of factors of economy, convenience and interpretability.

Test of Validity

- The extent to which the instrument measures what it claims to measure.
- For example, a test that is used to screen applicants for a job is valid if its scores are directly related to future job performance.

(i) Content validity

(ii) Face validity

(iii) Criterion validity

(iv) Construct validity

Content Validity

- Content validity pertains to the degree to which the instrument fully assesses or measures the construct of interest.
- It occurs when the experiment provides adequate coverage of the subject being studied.

Face Validity

- The term face validity has a similar meaning.
- However, face validity generally refers to "non-expert" judgements of individuals completing the instrument and/or executives who must approve its use.
- Respondents may refuse to cooperate or may fail to treat seriously measurements that appear irrelevant to them.
- Therefore, to the extent possible, researchers should strive for face validity.

Criterion Validity

- The criterion validity relates to our ability to predict some outcome or estimate the existence of some current condition.
- It reflects the success of measures used for some empirical estimating purpose.
- Any criterion must be judged based on four qualities:

  (i) Relevance

  (ii) Freedom from bias

  (iii) Reliability

  (iv) Availability

- A criterion is relevant if it is defined in terms we judge to be the proper measure.
- Freedom from bias is attained when the criterion gives each subject an equal opportunity to score well.
- A reliability criterion is stable and reproducible.
- Finally, the information specified by the criterion must be available. If it is not available, how much it will cost and how difficult it will be to secure?
- Criterion-related validity is expressed as the coefficient of correlation between test scores and some measure of future performance or between test scores and scores on another measure of known validity.

Construct Validity

- The construct validity is the most complex and abstract.
- In attempting to evaluate construct validity, we consider both the theory and the measuring instrument being used.
- A measure is said to process construct validity to the degree that it confirms to predicted correlations with other theoretical propositions.

Test of Reliability

- The reliability of a measure indicates the extent to which it is without bias (error-free) and thus ensures consistent measurement across time and across the various items in the instrument.
- A measure is reliable to the degree that it supplies consistent results.
- Reliability is necessary contributor to validity but is not a sufficient condition for validity.
- The reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept and thus helps to access the goodness of a measure.

Stability of Measures

- A measure is said to process stability if one can secure consistent results with repeated measurements of the same subject (respondent) with the same instrument.
- Two tests of stability are

  (i) test-retest reliability

  (ii) parallel-form reliability

Internal Consistency of Measures

- The internal consistency of measures is indicative of the homogeneity of items that measure the construct.
- The items should hang together as a set and be capable of independently measuring the same concept so that the subjects attach the same overall meaning to each of the items.
- This can be examined through the following two tests.

Inter-item Consistency Reliability

- Inter-item consistency reliability also simply known as "internal consistency" measures the degree to which different items measuring the same construct attain consistent results.
- Scores on different items designed to measuring the same construct should be highly correlated.
- The most popular test of inter-item consistency reliability is the Cronbach's coefficient of alpha.
- Cronbach’s alpha reliability : 0.6-0.7 indicates an acceptable level of reliability and 0.8 or greater a very good level.

Split-half Reliability

- Split-half reliability reflects the correlations between two halves of an instrument.
- The two halves can be created by splitting the items I several ways:

  (i) Odd and even numbered items

  (ii) First and second halves

  (iii) Randomly

- If the results of the correlation are high, the instrument is said to have high reliability in an internal consistency tells us there is homogeneity among the items.

Test of Practicality

- Economy consideration suggests that some trade-off is needed between ideal research project and that which the budget can afford.
- Convenience test suggests that measuring instrument should be easy to administer. One should give due attention to the proper layout of the instrument with clear instructions and coding.
- Interpretability consideration is more important when persons other than the designer of the instrument are to interpret the results.


Thank you
for your
attention

Questionnaire Design and Fieldwork Plan

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Learning Outcomes

After reading this chapter, student be able to:

- Understand the basic rules of questionnaire design
- Implement do’s and don’ts while designing the questionnaire
- Know how to validate the survey questionnaire
- Know how to per-test the questionnaire
- Understand how to plan for fieldwork for survey

Introduction

- The design of a questionnaire depends on the purpose of data collection.
- It depends on whether the researcher’s objective is to collect qualitative information or quantitative information.
- The qualitative information is needed if the study is more exploratory in nature for the purpose of better understanding of a given research problem or the generation of hypotheses on a subject.
- The quantitative information is needed if the study is more of conclusive in nature which requires testing of hypotheses that have been previously generated.

Exploratory questionnaires

- It may not be necessary to have formal questionnaire if the data to be collected is qualitative in nature and thus, does not require statistical validation.
- The researcher might prepare a brief guide, listing of important open-ended questions with appropriate probes/ prompts listed under each.

Formal standardized questionnaires

- If the researcher is interested to statistically analyse the data to test the hypotheses for a conclusive study, a formal standardized questionnaire must be designed.
- Such questionnaires are generally characterized by:

  - Prescribed wording and order of questions, to ensure that each respondent receives the same stimuli.

  - Prescribed definitions or explanations for each question, to ensure interviewers handle questions consistently and can answer respondents' requests for clarification if they occur.

  - Prescribed response format, to enable rapid completions of the questionnaire during the interviewing process.

Questionnaire Design for Business Research

- If the same study with exactly same hypotheses is being set for five different researchers, most likely each of them may come up with their own unique questionnaire.

- All the five questionnaires set by five researchers independently may differ widely in their choice of questions, wording, coding, sequencing, use of open-ended questions and so on.

There are a number of points that should be borne in mind.

- A questionnaire should be designed in such a way that it meets the objectives of the study set by the researcher.

- The questionnaire design should be such that it obtains the complete and accurate information as far as possible. The researcher must ensure that questions are asked in such a way that respondent fully understand the meaning of the questions and are not likely to refuse to answer or lie to interviewer.

- A well-designed questionnaire should ensure that it is easy for the respondents to respond to questions as well as easy for the interviewer to record the answer of the respondents.

- The questionnaire should be designed in such a way that it makes the interview concise and to the point, so as to make respondents remain interested till the end of the survey.

Defining the Target Population

- The researcher must define the target population about which he/she wishes to generalize from the sample data to be collected.

- Secondly, researchers have to draw up a sampling frame, method of sampling and the sample size.

- Thirdly, in designing the questionnaire we must take into account demographic information such as the age, education, etc. of the target respondents to ensure that the characteristics of the sample are as similar as the characteristics of the target population.

Language

- A questionnaire printed in English could be administered to the respondent in the local language he/she speaks by a trained interviewer who could translate each question on the spot.

Deciding on What to Ask

- The next obvious question is “What you are going to ask?

- In order to meet the objectives of the survey, it is important to decide what information one need to collect from the respondents.

- A number of questions might come in one’s mind related to the research problem that under investigation.

- But here the question is what questions should be asked and what questions should be avoided.

There are three potential types of information, researchers should be interested in:

- Information they are primarily interested in, that is, dependent variables.
- Information which might explain the dependent variables, that is, independent variables.
- Other factors related to both dependent and independent factors which may distort the results have to be adjusted for, that is, confounding variables.

Deciding How to Ask

❖ Questions can be asked in the close-ended form or open-ended form.

	Advantages	Disadvantages
Closed-ended	▪ Easy and quick to answer ▪ Easy to compare responses across the respondents ▪ Answers easier to analyse on computer ▪ Response choices make questions clearer ▪ Easy to replicate study	▪ Can put ideas in respondent's head ▪ Respondents can feel constrained/frustrated ▪ Many choices can be confusing ▪ Cannot tell if respondent misinterpreted the question ▪ Fine distinctions may be lost
Open-ended	▪ Permit unlimited number of answers ▪ Respondents can quality and clarify responses ▪ Can find the unanticipated answer ▪ Reveal respondents thinking process	▪ Answers can be irrelevant ▪ Inarticulate or forgetful respondents are at disadvantage ▪ Coding responses are subjective and tedious ▪ Requires more response time and effort

Wording of Individual Questions

- Use short and simple sentences
- Ask for only piece of information at a time
- Avoid negatives if possible
- Ask precise questions
- Questions must generate
- the required information
- Words must have the same meaning to all respondents
- Avoid leading questions
- Avoid hypothetical questions
- Do not overtax the respondent's memory
- Ensure those you ask have the necessary knowledge
- Level of details
- Sensitive issues
- Minimize bias

Sequencing of Questions

- Opening questions
- Question flow
- Question variety

The following points should be considered while sequencing questions in the questionnaire.

- Put the most important items in the first half of the questionnaire.
- Do not start with awkward or embarrassing questions.
- Start with easy and non-threatening questions.
- Go from the general to the particular.
- Go from factual to abstract questions.
- Go from closed to open questions.
- Leave demographic and personal questions until last.

Length of Questionnaire

- There are no universal agreements about the optimal length of questionnaires.

- It probably depends on a number of factors such as the number of objectives in the study,
  - ✓ type of respondents (whether target respondents are consumers, managers, students or kids)
  - ✓ type of survey (whether it is a personal interview, postal survey or internet survey).

- As a rule of thumb, ask only necessary questions and avoid unnecessary one.

Ease of Recording

Questionnaire design should ensure.

- It is easy to carry, visible in different kinds of light.
- The distance between different answer categories should be sufficient.
- There is no confusion or mistake while placing a tick over the actual response or a given question.

Coding

- If the questionnaire is coded before doing the fieldwork (as most of the questionnaire are likely to be these days), it must be ensured that the fieldworkers know where to mark the answer - on the code or actual answer choice.

Analysis Required

- It is important to plan for analysis well before designing the questionnaire.

- Similarly, if you plan to analyse the factor analysis to empirically find out important dimensions of a theoretical construct, the questions need to be measured on a rating scale.


Thank you
for your
attention

Questionnaire Design and Fieldwork Plan

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Learning Outcomes

After reading this chapter, student be able to:

- Understand the basic rules of questionnaire design
- Implement do’s and don’ts while designing the questionnaire
- Know how to validate the survey questionnaire
- Know how to per-test the questionnaire
- Understand how to plan for fieldwork for survey

General Appearance of a Questionnaire

- The physical appearance of a questionnaire can have a significant effect upon both the quantity and quality of survey data obtained.

- The appearance of questionnaire has an impact on respondents' motivation in giving response.

Introduction to Respondents

- It is important to have the covering letter if the questionnaire is to be mailed or distributed to respondent.

- The purpose of such a letter is to introduce the respondents with the objectives of the survey in order to encourage them to actively participate in the survey.

Instructions

- Interviewer instructions should be placed alongside the questions to which they pertain.

- Instructions on where the interviewers should probe for more information or how replies should be recorded are placed after the question.

Concluding Questionnaire with an Open-ended Question and Thanks

- The questions that are of special importance should be placed in the earlier part of the questionnaire and the sensitive questions towards the end of the questionnaire.

- It is always good to end the questionnaire with an open-ended question to get the free opinion on the topic.

- At the end of the questionnaire do not forget to "thank" the respondent once again for their valuable time spent on completing the survey.

Pre-testing Questionnaire

- Pre-testing the questionnaire is an essential step before its completion.

- The purpose of the pre-test is to check question wording, and to obtain information on open-ended questions with a view to design a multiple choice format in the final questionnaire.

The purpose of pre-testing the questionnaire is to determine:

- Wording of the questions are correct to convey the same meaning to all the respondents.
- Whether the questions have been placed in the right sequence.
- Whether the questions are clearly understood by all classes of respondents.
- Whether additional questions are needed or whether some questions should be eliminated.
- Whether the instructions to interviewers are clear and adequate.

Design of Fieldwork

- Careful planning is required for prompt receipt of survey questionnaires to the different areas covered in the data collection undertaking.

- A clear plan for data collection should be developed by the researcher.

Fieldwork Plan

- Fieldwork plan is clearly linked to the sampling plan.
- Once the sampling area (cities, town, etc.) and the sample size are determined for each, the next step is to plan on the following:
  - Who will do the fieldwork?
  - When should the fieldwork start?
  - How long should the fieldwork be carried on?
- First, you need to plan who will do the fieldwork.
- The second question is when the data collection should be carried out.

❖ For the third issue, i.e. for the time requirement for data collection, one needs to consider the following points.

Step 1: Calculate the:

- time required to reach the study area(s),
- time required to locate the study units (persons, groups or records), and
- number of visits required per study unit. In the longitudinal studies, the survey needs to be carried out from time to time, whereas in the cross- sectional study, the survey is done once forever.

Step 2: Calculate the number of interviews that can be carried out per field-worker per day.

Step 3: Calculate the number of days needed to carry out the interviews.


Thank you
for your
attention

Efficient Data Collection and Exportation from Google Forms

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Google

A multinational technology company that specializes in internet-elated services and products.

Google Apps

Gmail, Drive, Docs, Sheets, Slides, Meet, Classroom, etc.

Google Form

An online tool that allows user to create and distribute surveys, quizzes and forms.

Export Data to Excel

Export the collected data for further analysis or processing in Excel.



01

Creating a form

02

Collecting data

03

Data analysis

04

Advanced feature

Creating A Google Form

- Start by deciding what type of questions you want to ask.
- Go to forms.google and log in with Google account.
- Create a new form and sometimes, the sample templates are useful.
- Customize the form by adding questions and selecting the appropriate respond types.
- Customize the design and appearance of the form using themes and colors.
- Lastly, share the form to the respondents.



Collecting Data



Once Google Form is live,

- Start collecting responses.
- Be able to share the form : via email, social media, direct link or embedding it in a website.
- Once respondents start filling out the form, the response will be automatically collected and stored.
- View responses in real-time
- Send out reminders to those who haven't responded yet.
- Export data into different formats, like CSV or Excel , to further analyze the data.

Basic Data Analysis on Google Form

Summary of Responses

Be able to view a summary of responses.

- Provides an overview of response counts, average values and charts for multiple choice questions.
- Automatically generate bar chart and pie chart.
- Give a quick snapshot of the data.



Basic Data Analysis on Google Sheets



Downloading Data

- Export the collected data from Google Forms to a Google Sheets spreadsheet.
- Provide more flexibility for in-depth data analysis using spreadsheet functions, formulas and advanced charting options available in Google Sheets.

Third-Party Tools

If you require more advance data analysis

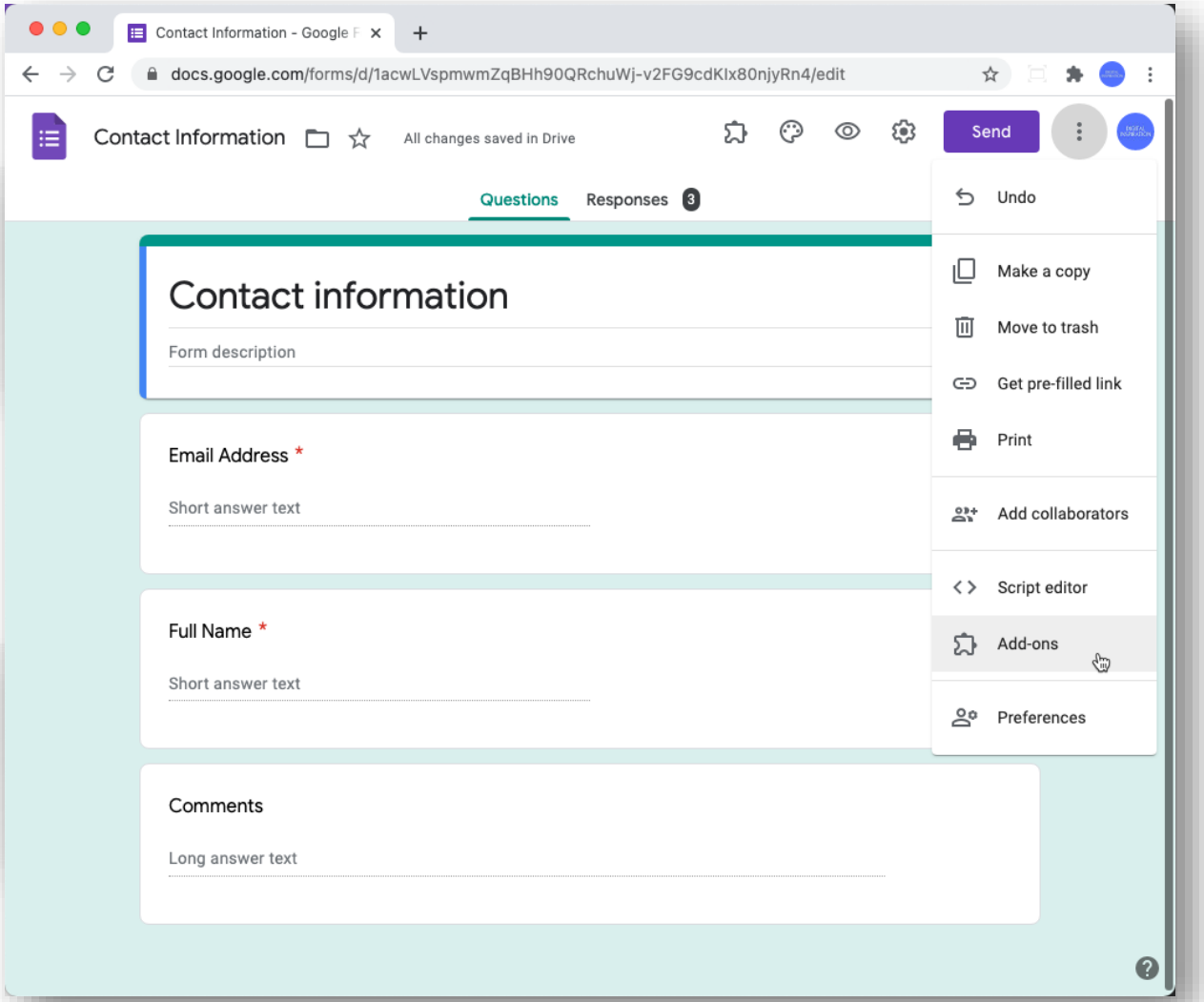
- Export the data from Google Sheets to other data analysis tools like Microsoft Excel, Google Data Studio or Statistical analysis software such as SPSS, R, Python, Eviews and so on.
- These tools offer extensive capabilities for data manipulation, visualization and advanced statistical analysis.

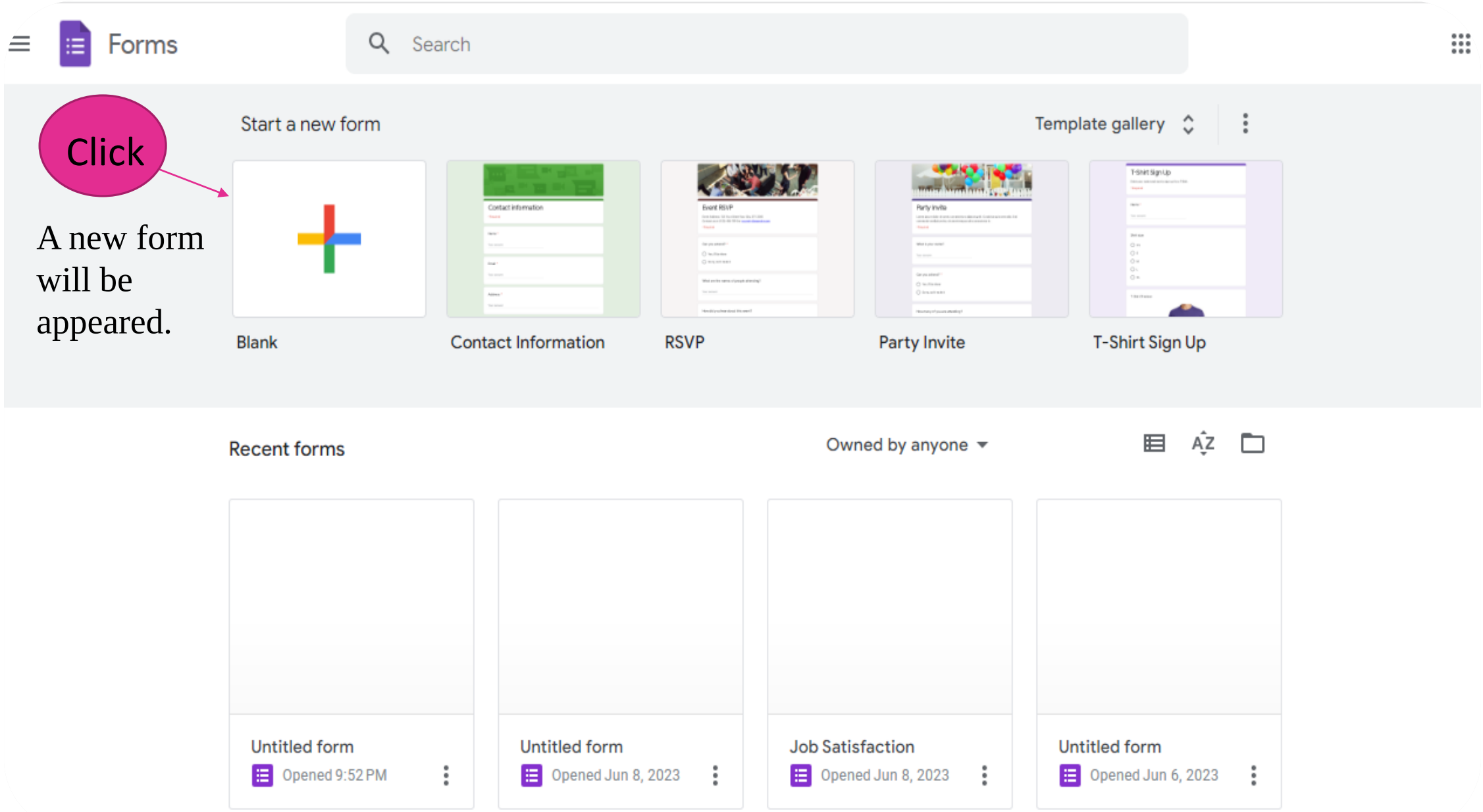
Basic Data Analysis on Google Sheets

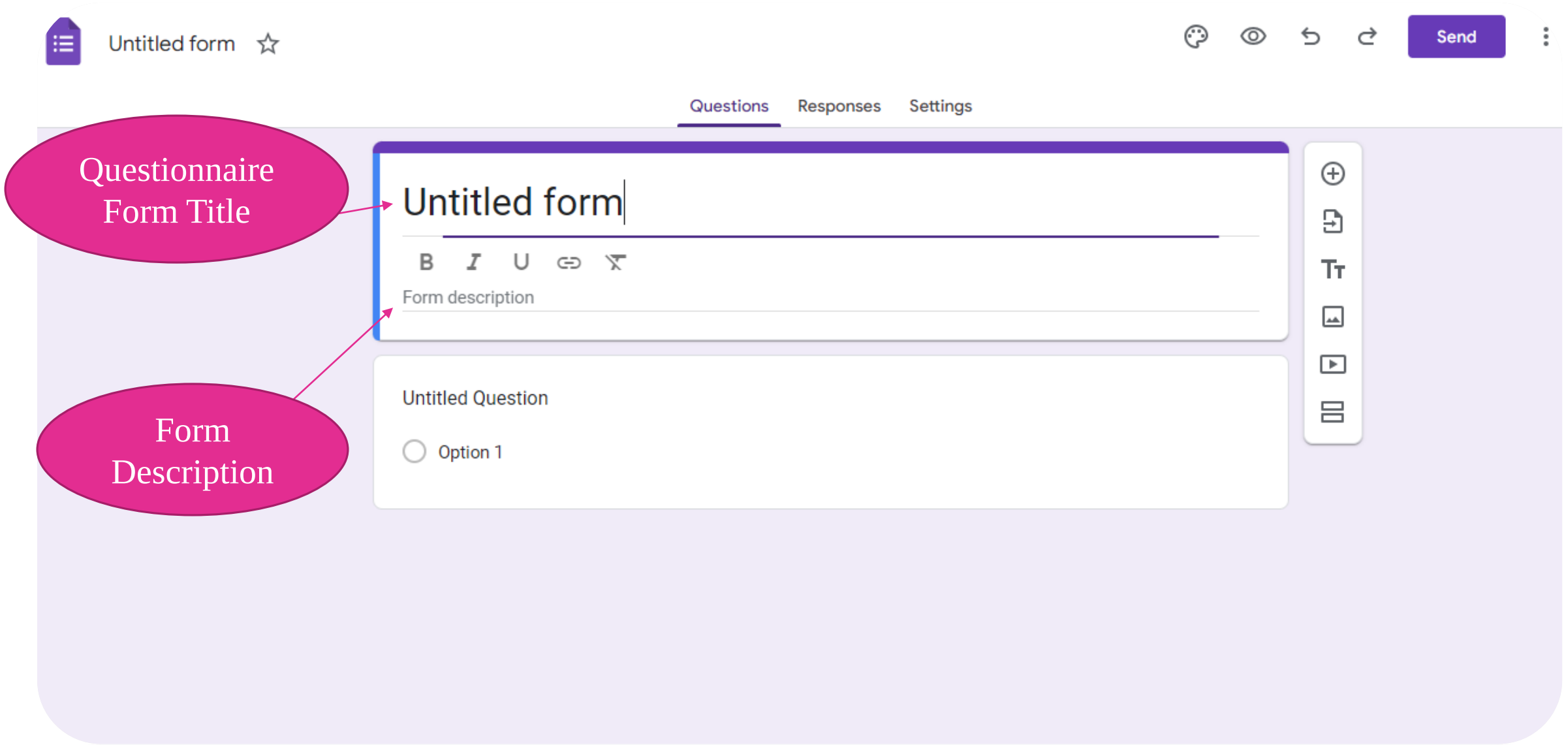


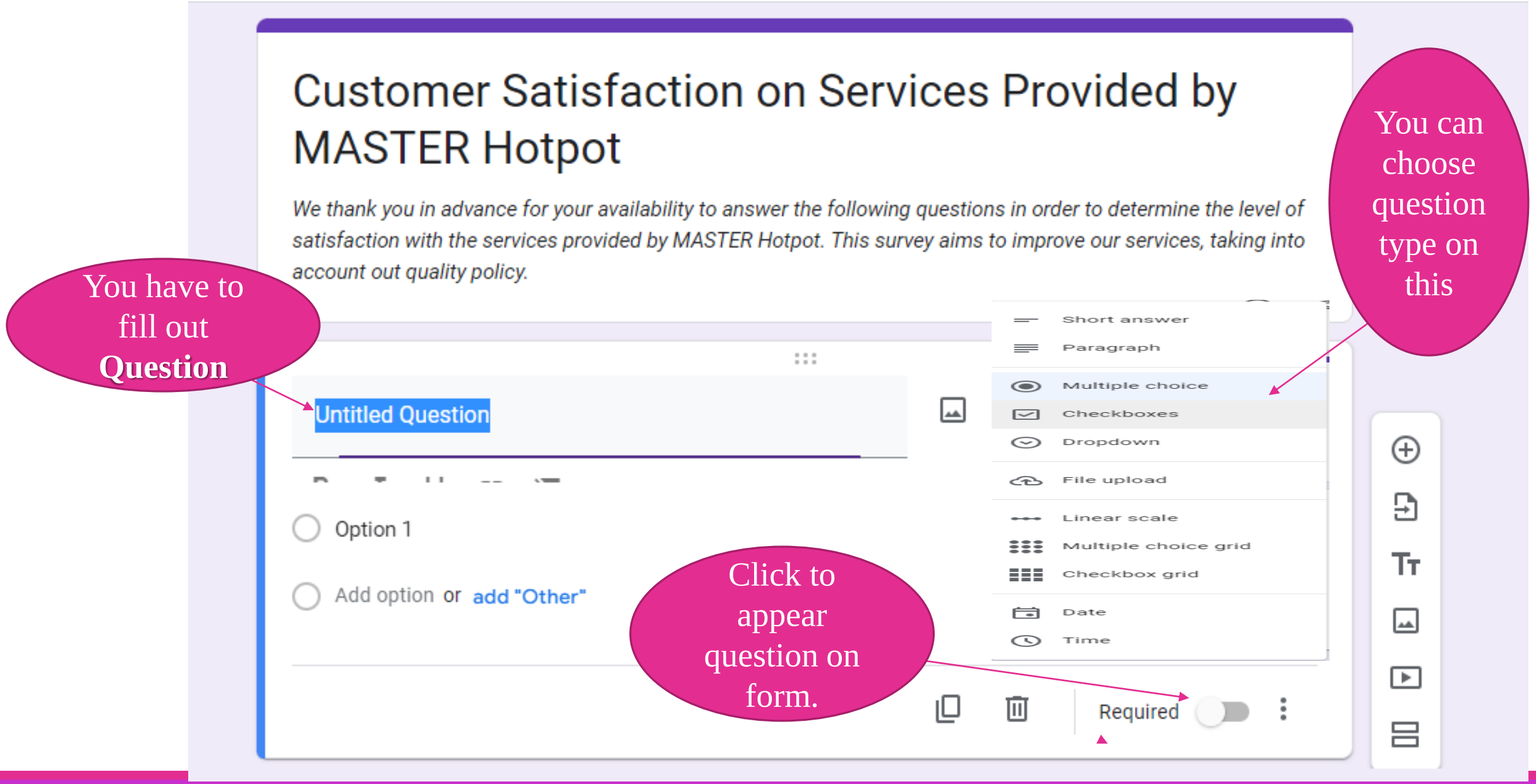
1 Google Form

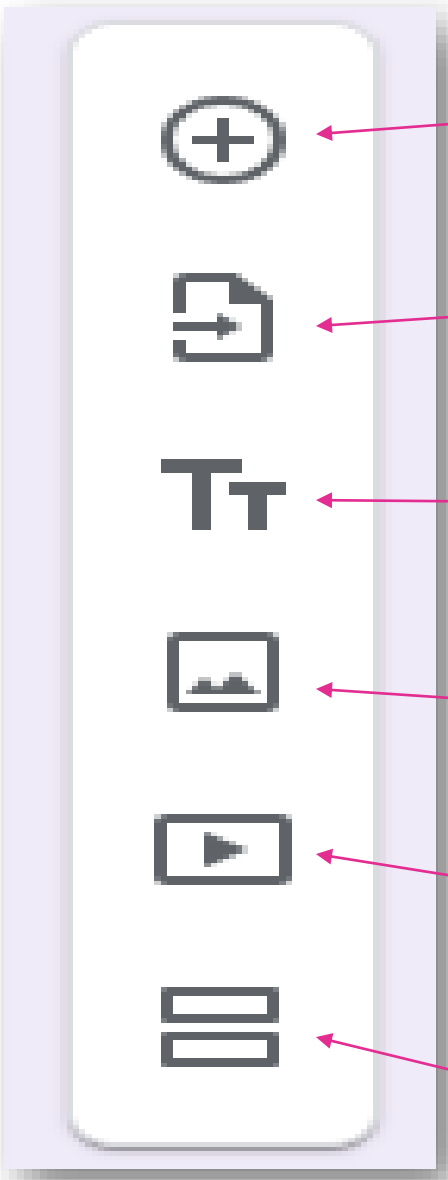
2 Start a new form or sample templates




Forms
Search
Click
A new form will be appeared.
Start a new form
Template gallery
Blank
Contact Information
RSVP
Party Invite
T-Shirt Sign Up
Recent forms
Owned by anyone
Untitled form
Opened 9:52 PM
Untitled form
Opened Jun 8, 2023
Job Satisfaction
Opened Jun 8, 2023
Untitled form
Opened Jun 6, 2023


Untitled form
Send
Questions
Responses
Settings
Questionnaire Form Title
Untitled form
Form description
Form Description
Untitled Question
Option 1


Customer Satisfaction on Services Provided by MASTER Hotpot
We thank you in advance for your availability to answer the following questions in order to determine the level of satisfaction with the services provided by MASTER Hotpot. This survey aims to improve our services, taking into account out quality policy.
You have to fill out Question
You can choose question type on this
Untitled Question
Option 1
Add option or add "Other"
Short answer
Paragraph
Multiple choice
Checkboxes
Dropdown
File upload
Linear scale
Multiple choice grid
Checkbox grid
Date
Time
Click to appear question on form.
Required



Add question

Import question

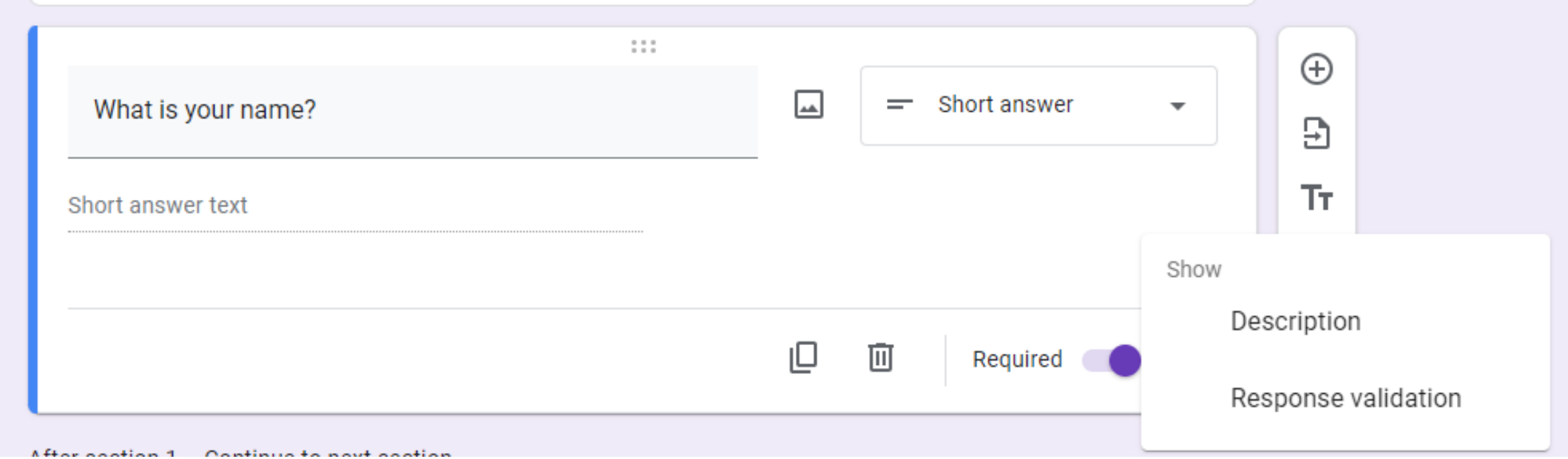
Add title and description

Add image

Add video

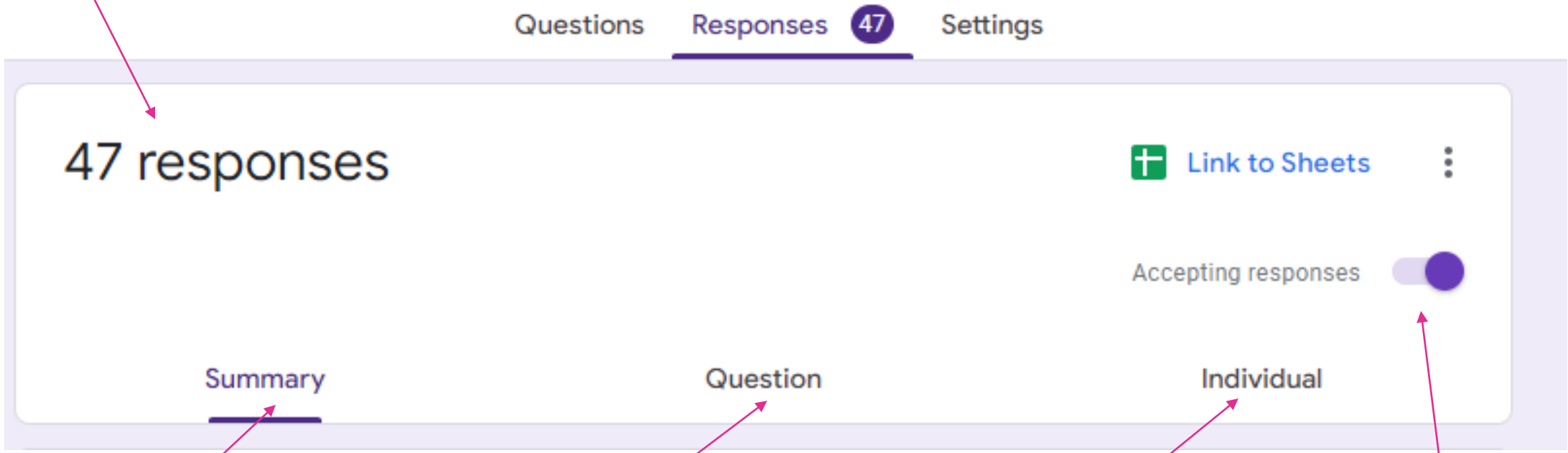
Add Section

Question with Short Answer



Number of Responses

You can check how many responses there are.

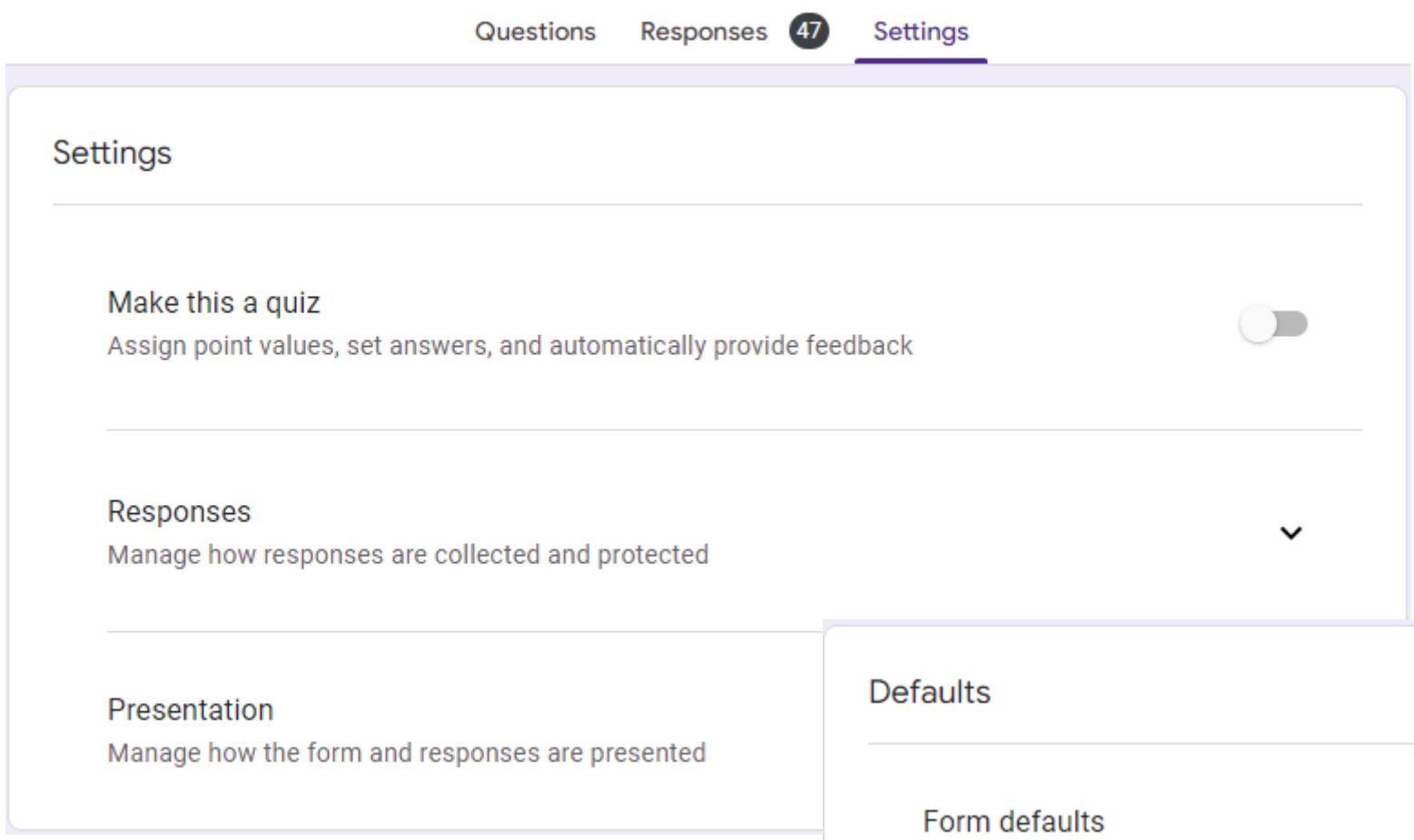


Data presentation by charts

Show the number of response on each category

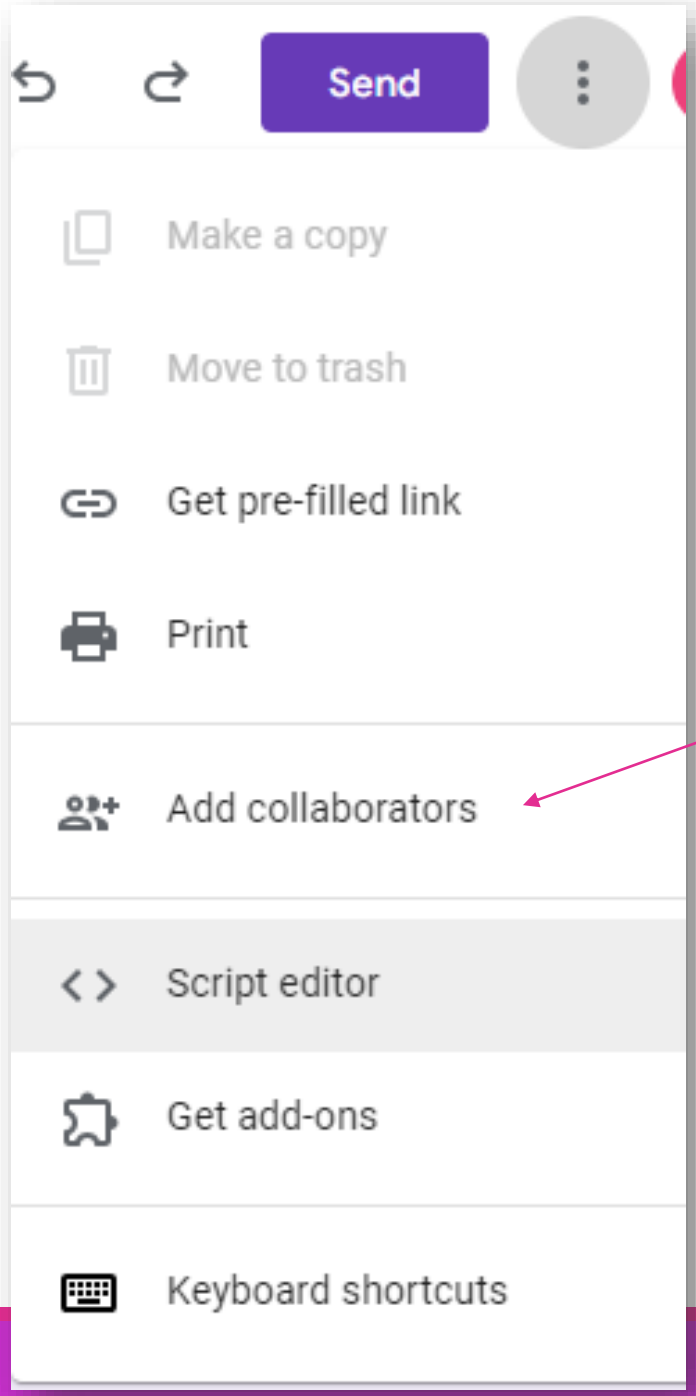
Provide Individual response

If you stop responding, you can switch off.

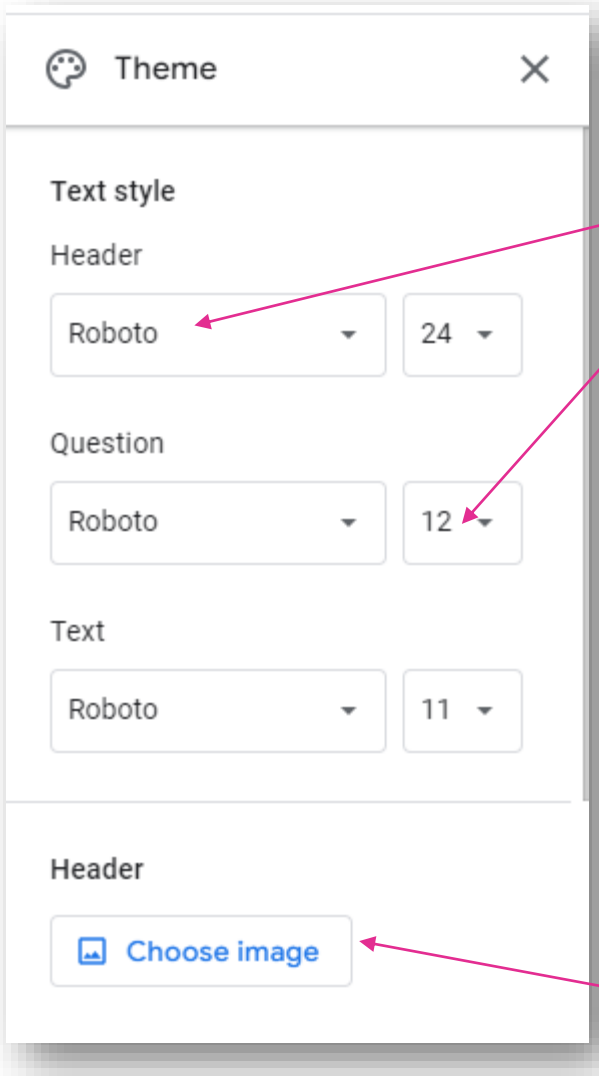


Setting

You can adjust your from in this section.

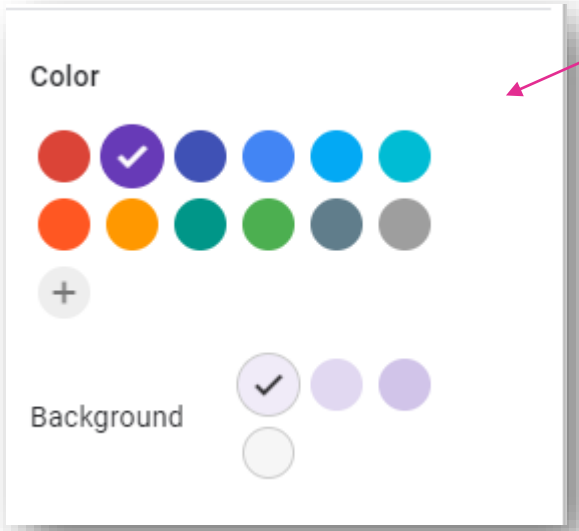


You can add collaborators who can also edit the form and view the response.

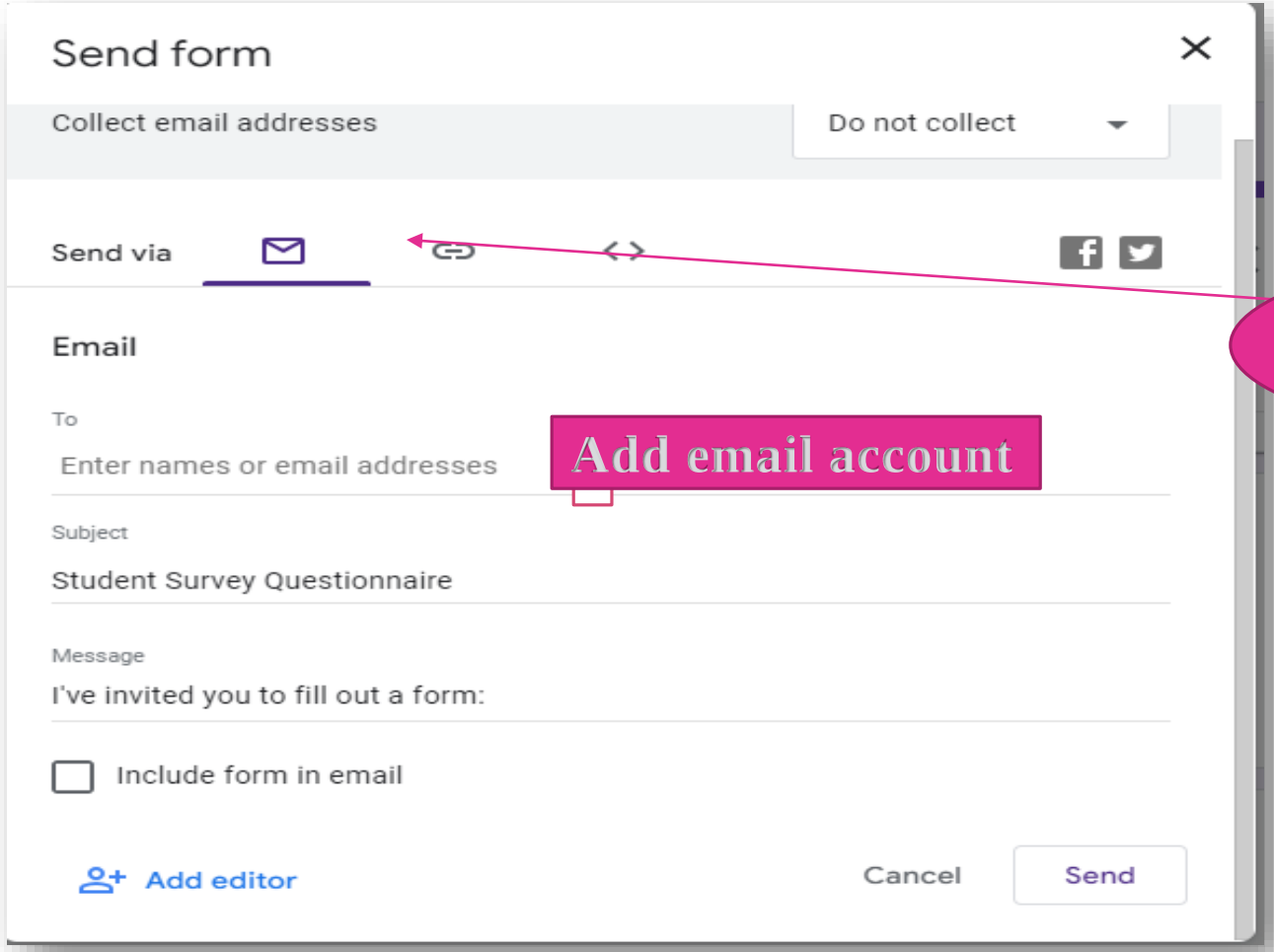


You can change text style and text size.

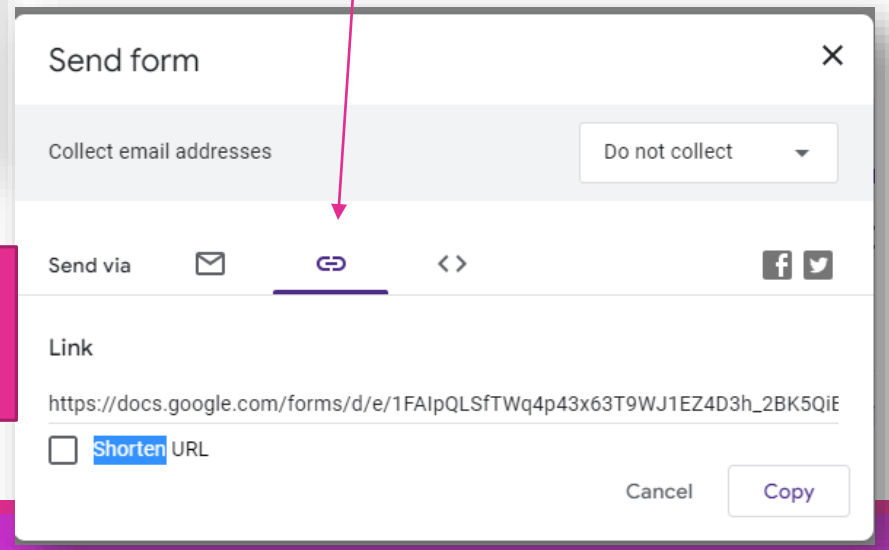
You can choose any color for the form.



You can choose the image as the header background.

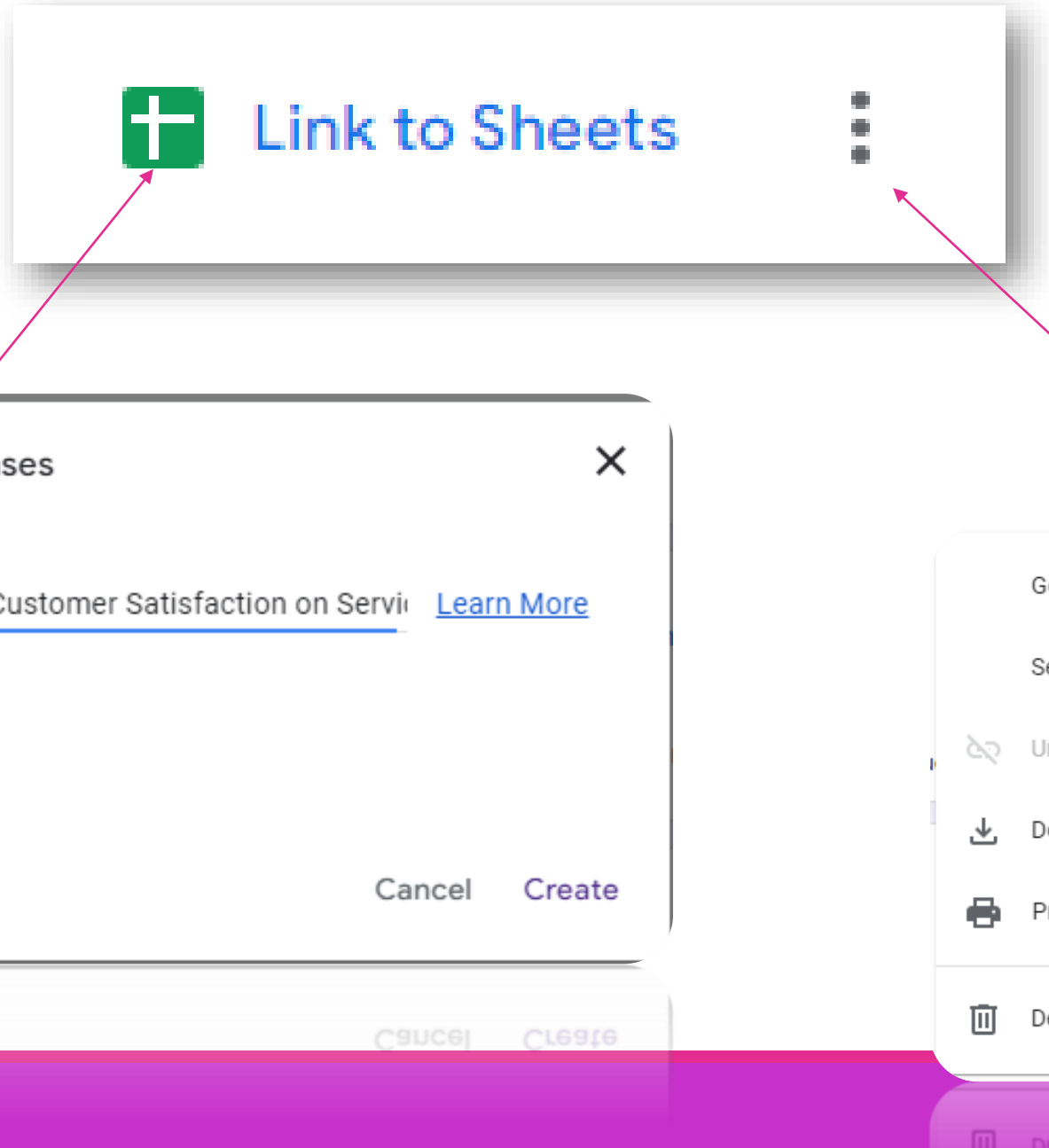


You can send this form via email or using link



Copy either long link or short link

Send


Link to Sheets
Select destination for responses
Create a new spreadsheet
Customer Satisfaction on Servi
Learn More
Select existing spreadsheet
Cancel
Create
Get email notifications for new responses
Select destination for responses
Unlink form
Download responses (.csv)
Print all responses
Delete all responses

Questionnaire Design

- Begin with short, clear instructions.
- State the survey purpose.
- Assure anonymity.
- Instruct on how to submit the completed survey.

Questionnaire Design

- Break survey into naturally occurring sections.
- Let respondents bypass sections that are not applicable (e.g., “if you answered no to question 7, skip directly to Question 15”).
- Pretest and revise as needed.
- Keep as short as possible.

Questionnaire Design

Type of Question	*Example*
Open-ended question	Briefly describe your job goals.
Fill-in-the-blank	How many times did you attend formal religious services during the last year? _____ times
Check boxes	Which of these statistics packages have you ever used? ❑ SAS ❑ Visual Statistics ❑ SPSS ❑ MegaStat ❑ Systat ❑ Minitab

Questionnaire Design

Type of Question	*Example*
Ranked choices	“Please evaluate your dining experience”

	Excellent	Good	Fair	Poor
Food	❑	❑	❑	❑
Service	❑	❑	❑	❑
Ambiance	❑	❑	❑	❑
Cleanliness	❑	❑	❑	❑
Overall	❑	❑	❑	❑

Questionnaire Design

Type of Question	*Example*
Pictograms	"What do you think of the President's economic policies?" (circle one) 
Likert scale	Statistics is a difficult subject. Strongly Agree / Slightly Agree / Neither Agree Nor Disagree / Slightly Disagree / Strongly Disagree ❑ ❑ ❑ ❑ ❑


THANK
YOU

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

PGDRS-111

DATA STRUCTURE AND DATA COLLECTION METHODS

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Chapter 9
Sampling Design: Theory and Practice

Learning Outcomes

After reading this chapter, student should be able to:

- ❑ Understand why you need to sample the population
- ❑ Know the basic terminology
- ❑ Understand the differences between probability and non-probability sampling
- ❑ Apply the appropriate sampling technique
- ❑ Determine the sample size
- ❑ Understand the factors that could affect the sample size in any study
- ❑ Understand the different types of errors in research

Introduction

- The concept of sampling is used in our day to day life.
- Sampling may be defined as the selection of some part of the population on the basis of which a judgement or inference about the entire population is made.
- In most of the research work and surveys, the usual approach is to make generalizations or to draw inferences based on samples about the parameters of the population from which the samples are taken.
- So the sample should be drawn in such a way that it is true represent of the entire population.
- The process of sample selection is called sample design and the survey conducted on the basis of sample is described as sample survey.

Why Sampling?

- The question is "Why sample the population?" The following points justify why we need to choose the sample rather than going for complete census of the target population.
  - The population is dynamic, i.e. the component of the population could change over time. Thus, it is practically impossible to check all items in the population.
  - The cost of studying the entire population could be very high. A sample study is usually less expensive than a census study.
  - Contacting the whole population would often be time consuming. Sampling can save time as the results can be produced at a relatively faster speed.
- Sampling remains the only way when population contains infinitely many members or when the experiment involves the destruction of the items under study.
- Sampling usually enables to estimate the sampling errors and assists in obtaining information concerning some characteristics of the population.

Defining Basic Terminology

- ❑ Population, Element and Population Size
- ❑ Sample, Subject and Sample Size
- ❑ Parameter and Statistics
- ❑ Sampling Frame

Population, Element and Population Size

- Population refers to the entire group of people, events or things of interest that researcher wishes to investigate.
- Each member of the population is known as element.
- The total number of elements in the population is known as population size and it is denoted by “N”.

Sample, Subject and Sample Size

- The sample is the subset of the population.
- Each member of the sample is known as subject.
- The total number of subjects in the sample is known as sample size and it is denoted by “n”.

Parameter and Statistics

- The characteristics of the population are known as parameters whereas the characteristics of the sample are known as statistics.
- Statistic is used to estimate the value of the parameter.
- Note that the value of statistic changes from one sample to the next which leads to a study of the sampling distribution of statistic.

Sampling Frame

- It is a complete listing of the population of interest from which the sample is drawn.
- All members of the sampling frame have a probability of being selected.
- Without some form of sample frame, a random sample of a population, other than an extremely small population, is impossible.


Thank you
for your
attention

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

PGDRS-111

DATA STRUCTURE AND DATA COLLECTION METHODS

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

Chapter 9
Sampling Design: Theory and Practice

Learning Outcomes

After reading this chapter, student should be able to:

- ❑ Understand why you need to sample the population
- ❑ Know the basic terminology
- ❑ Understand the differences between probability and non-probability sampling
- ❑ Apply the appropriate sampling technique
- ❑ Determine the sample size
- ❑ Understand the factors that could affect the sample size in any study
- ❑ Understand the different types of errors in research

Sampling Techniques

Probability Sampling Techniques

- In probability sampling technique, a sample is being selected using random selection so that each element of the population has a known chance of being selected.

- It is generally assumed that a representative sample is more likely to be the outcome when this method of selection from the target population is employed.

- Thus, findings based on probability sampling can be generalized to the target population with a specified level of confidence.

Simple random sampling

- The simple random sampling is the most basic form of probability sampling.
- Here the sample is drawn from the target population in such a way that each and every member of the population has an equal and known chance of being the subject of the sample.

The procedure for drawing large sample involves the following steps:

- Sequentially assign a unique identification number to each element of the population.
- Use a random number generation (such as lottery method) to identify the appropriate element to be the part of sample.
- Ensure that no element is selected more than once.

N=300 employees

n=300*20%=60 employees

The sample number are

56	114	186	111	142	136	97	188	79	38	58	273	77	225	139
53	241	163	300	159	152	124	141	79	4	129	239	186	134	249
173	102	188	142	147	45	198	300	164	98	134	20	9	270	266
102	280	233	103	23	59	4	269	175	80	216	234	33	43	157

Stratified sampling

- In stratified random sampling, the population is divided into different subgroups known as strata on the basis of some criteria.
- Then the method of simple random sampling is used to draw the sample from each stratum.
- Now the sample can be drawn from each stratum either proportionately or disproportionately.
- In proportionate stratified random sampling, the sample taken from each stratum will be equal to 20% of the population size of the strata as shown in column 3.

- In disproportionate stratified random sampling, the percentage of population taken as the sample for different strata is not the same.

- Disproportionate sampling decisions are made either when some stratum or strata are too small or too large, or when there is more variability suspected within a particular stratum.

- Disproportionate sampling is also sometimes done when it is easier, simple, and less expensive to collect data from one or more strata than others.

N=885 employees

n=885*20%=177 empployees

JL	Number of employee	Prp sample	Nonprop sample
TM	15	3	
MM	40	8	
LM	60	12	
S	140	28	
C	600	120	
S	30	6	

Systematic Sampling

- Systematic sampling is very similar to random sampling, and is easier in practice.
- Once the sample size is decided, we divide the total population into n parts, where ‘n’ is the sample size required.
- From the first part of sampling units, we pick up one at random.
- We then pick up every $(N/n)^{th}$ item from the remaining parts.

N=300 households

n=15 households

k=N/n=300/15=20

20th every households

First no.= 7

The sample number are

7,27,47,67,87,.....

Cluster Sampling

- A cluster is a group of sampling units or elements which can be identified or listed and a sample of which can be chosen.
- In simple random sampling, elements are randomly selected from the entire population of study.
- In cluster sampling, population is divided into groups of elements with some groups randomly selected for the study.
- However, it is different from stratified sampling.

- The selection of a sample of clusters to provide a sample of population units is called cluster sampling.
- The two primary reasons cluster sampling is employed for sample surveys of human populations in large geographical regions are feasibility and economy.
- Cluster sampling is often the only feasible method of probability based sampling because the sampling frames for the target populations are lists of clusters.
- Further, it is economical compared to simple random sampling as the subjects of the sample are selected from the randomly selected strata rather than from the entire population.
- However, the statistical efficiency of cluster sampling is usually lower than simple random sampling.

N=500 Households

Ward 1,2,3,4,5

n=100 households

The sample number are

Ward	N	n
1	50	50
2	100	
3	50	50
4	100	
5	100	

Multistage cluster sampling

- In single-stage cluster sampling, the population is divided into convenient clusters and we randomly choose the required number of clusters as sample subjects and investigate all the elements in each of the randomly chosen cluster.

- Cluster sampling can also be done in several stages and it is known as multistage cluster sampling.

Non-probability Sampling

- Some of the non-probability sampling techniques are commonly used explicitly in cases where it is not possible to use the probability sampling.
- The major difference is that in non-probability techniques, the extent of bias in selecting the sample is not known.
- This make difficult to say anything about the representativeness or accuracy of the sample.
- Nevertheless, if done carefully, some of these are good approximation of probability sampling.

Convenient Sampling

- It refers to the collection of information from members of the population who are conveniently available to provide it.
- It involves picking up any available set of respondents convenient for the researcher to use.
- The convenience sampling is very often used by researchers in order to cover the large number of survey quickly and cost effectively.
- However, it suffers from selection bias because the individuals surveyed are often different from the target population and thus may not be a true representative of the population for the study in hand.

Judgement Sampling

- A judgement sample, sometimes referred to as a purposive sample, involves selecting elements in the sample for a specific purpose.

- It involves the choice of subjects who are most advantageously placed or in the best position to provide the information required.

- Judgement sample might be a group of experts with knowledge about a particular problem or issue.

Quota sampling

- Quota sampling is a non-probability sampling technique, wherein certain groups are adequately represented in the study through the assignment of a quota based on known characteristics.
- Quota sampling could be either proportional or disproportional.
- In proportional quota sampling, we represent the major characteristics of the population by sampling a proportional amount of each.
- Disproportional quota sampling is a bit less restrictive. In this method, we specify the minimum number of sampled units we want in each category.
- This method is typically used to assure that smaller groups are adequately represented in the sample.

Determining the Sample Size

- Determining sample size is a very important issue because samples that are too large may waste time, resources and money, while samples that are too small may lead to inaccurate results.

- To determine the sample size three pieces of information are required. They are:

  i. the degree of confidence necessary to estimate true value,

  ii. the precision of the estimate and

  iii. the amount of true variability present in the data.

Level of Precision

- Precision refers to how close our estimate is to the true population characteristics.
- We estimate the population parameter to fall within a range based on the sample estimate.
- The narrower this interval, greater is the precision.
- 995000-1050000
- 990000-1100000
- That is, we would now estimate the population mean to lie within narrower range, which in turn means that we would now estimate with greater precision.

Variability in Data

- Precision is a function of the range of variability in the sampling distribution of the sample mean.
- The smaller the dispersion or variability, the greater the probabilities that the sample mean will be closer to the population mean.
- However, there is no need to take several different samples to estimate the variability.
- This variability is known as standard error($S_{\bar{X}}$) and is calculated as:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Two points to be noted here.

(i) In order to reduce the standard error ($S_{\bar{X}}$), one needs to increase the sample size at a given standard deviation of the sample (S).

(ii) Smaller the variation in population, the smaller the standard error($S_{\bar{X}}$), which in turn implies that sample size need not be large.

- We need the greater precision if we want our sample results to closely reflect the characteristics of the populations.

- The greater the precision required, the larger is the sample size needed, particularly when the variability in the population is large.

Level of Confidence

- The confidence or risk level is based on ideas encompassed under the Central Limit Theorem.

- The key idea encompassed in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value.

- In a normal distribution, approximately 95% of the sample values are within two standard deviations (2σ) of the true population mean (μ).

Sample Data, Precision and Level of Confidence in Estimation

- Precision and confidence play vital role in sampling as we use sample data to draw inferences about the entire population.

- Because the point estimate provides no measure of possible error, to be fair we do an interval estimation to ensure a relatively accurate estimation of the population parameters.

- The standard error $S_{\bar{X}}$ and the level of confidence determine the width of the interval and calculated by using the following formula.

- For 90% level of confidence, the Z value is 1.65.
- For 95% level of confidence, the Z value is 1.96.
- For 99% level of confidence, the Z value is 2.58.

Factors Affecting the Decision on Sample Size

Roscoe(1975) proposes the following rules of thumb for determining the sample size:

- Sample size larger than 30 and less than 500 are appropriate for most research.
- Where samples are to be broken into sub-samples such as male/female, Malay/Chinese/Indian, etc., a minimum sample size of 30 for each category is necessary.
- In multivariable data analysis (including multiple regression analysis), the sample size should be several times (preferably 10 times or more) as large as the number of variables in the study.
- For simple experimental research is possible with samples as small as 10 to 20 in size.

Types of Error in Research

Sampling Error

- This is the error which occurs due to the selection of some units and non-selection of other units into the sample.
- It is controllable if the selection of the sample is done in random, unbiased way.
- In other worlds, if the probability sampling is used, it is possible to control this error.
- In general, sampling error reduces as the sample size increases.

Non-sampling Error

- This is the effect of various errors in doing the study, by the interviewers, data entry operator or the researcher himself.

- Handling a large quantity of data is not an easy job, and errors may creep in at any stage of research.

- The data entry person may interchange the column of yes and no responses while entering and compiling the data, or the interviewer may cheat by not filling up the questionnaire in the field and instead, fudge the data.

The **total error** in research is the sum of above two errors.

Total Error = Sampling Error + Non-sampling Error

- The sampling error can be estimated in the case of probability sampling, but not in the case of non-probability sampling.
- Non-probability sampling can be controlled through hiring better field workers, qualified data entry persons, and good control procedures at every stage of research.


Thank you
for your
attention

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF APPLIED STATISTICS

PGDRS-111

DATA STRUCTURE AND DATA COLLECTION METHODS

Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

ORGANIZATIONAL RESEARCH: DETERMINING APPROPRIATE

SAMPLE SIZE IN SURVEY RESEARCH

James e. bartlett

Formula for the Minimum Sample Size Needed for an Interval Estimate of the Population Mean

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

where E is the margin of error. If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n.

EXAMPLE 7–4 Automobile Thefts

A sociologist wishes to estimate the average number of automobile thefts in a large city per day within 2 automobiles. He wishes to be 99% confident, and from a previous study the standard deviation was found to be 4.2. How many days should he select to survey?

SOLUTION

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.58$ and $E = 2$. Substitute in the formula

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.58)(4.2)}{2}\right]^2 = 29.35$$

Round the value up to 30. Therefore, to be 99% confident that the estimate is within 2 automobiles of the true mean, the sociologist needs to sample the thefts for at least 30 days.

Sample Size for Proportions

To find the sample size needed to determine a confidence interval about a proportion, use this formula:

Formula for Minimum Sample Size Needed for Interval Estimate of a Population Proportion

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

If necessary, round up to obtain a whole number.

This formula can be found by solving the margin of error value for n in the formula

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

EXAMPLE 7–11 Land Line Phones

A researcher wishes to estimate, with 95% confidence, the proportion of people who did not have a land line phone. A previous study shows that 40% of those interviewed did not have a land line phone. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

SOLUTION

Since $z_{\alpha/2} = 1.96$, $E = 0.02$, $\hat{p} = 0.40$, and $\hat{q} = 0.60$, then

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.40)(0.60)\left(\frac{1.96}{0.02}\right)^2 = 2304.96$$

which, when rounded up, is 2305 people to interview. So the researcher must interview 2305 people.

95%
Alpha=5%=0.05
$Z\alpha/2=Z_{0.05/2}=Z_{0.025}=1.96$
$Z_{0.05}=1.645$

99%
α=1%=0.01
$Z\alpha/2=Z_{0.01/2}=Z_{0.005}=2.58$

EXAMPLE 7–12 Home Computers

In Example 7–11 assume that no previous study was done. Find the minimum sample size necessary to be accurate within 2% of the true population.

SOLUTION

Here we do not know the values of $\hat{p}$ and $\hat{q}$. So we use $\hat{p} = 0.5$ and $\hat{q} = 0.5$.

$$E = 0.02 \qquad \text{and} \qquad z_{\alpha/2} = 1.96$$

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

$$= (0.5)(0.5)\left(\frac{1.96}{0.02}\right)^2$$

$$= 2401$$

Hence, 2401 people must be interviewed when $\hat{p}$ is unknown. This is 96 more people than needed if $\hat{p}$ is known.

In determining the sample size, the size of the population is irrelevant. Only the degree of confidence and the margin of error are necessary to make the determination.

Activat

Basic Sample Size Determination for Mean

95% confidence interval, Z=1.96

S=0.75, E=5% or less than 5%

$n=Z^2 S^2/ E^2 =(1.96)^2 *(0.75)^2/(0.05)^2 =865$

s= estimate of standard deviation =0.75

d= acceptable margin of error =0.05 or 0.03

Basic Sample Size Determination for Proportion

$$n = Z^2\ p\ q\ /\ E^2 = (1.96)^2 * 0.5 * 0.5\ /(0.05)^2 = 384$$

Z = 1.96

p = maximum possible proportion (0.5)

q= 1- maximum possible proportion= 1-(0.5) =0.5

E = acceptable margin of error for proportion = 0.05

Assuming a response rate of 65%, the required minimum sample size should be calculated base on the following:

Response rate = 65%.

Non response rate=35%

Sample size adjusted for response rate=384/0.65=591

Therefore, required minimum sample size of 591 should be used.

Sample Size Determination Table

Table (1) Table for Determining Minimum Returned Sample Size for a Given Population Size for Continuous and Categorical Data

Population Size	Sample size					
	Continuous data (margin of error=.03)			Categorical data (margin of error=.05)		
	alpha=.10 t =1.65	alpha=.05 t =1.96	alpha=.01 t =2.58	p =.50 t =1.65	p =.50 t =1.96	p =.50 t =2.58
100	46	55	68	74	80	87
200	59	75	102	116	132	154
300	65	85	123	143	169	207
400	69	92	137	162	196	250
500	72	96	147	176	218	286

Table (2) Minimum Number of Regressors Allowed for Sampling Example

Sample size for:	Maximum number of regressors if ratio is:	
	5 to 1	10 to 1
Continuous data: $\underline{n}$ = 111	22	11
Categorical data: $\underline{n}$ = 313	62	31

The sales manager of an electrical outlet wants to investigate the level of satisfaction among Customers who purchased the refrigerators from the outlet for the past 12 months. From the sales records of 480 customers the manager selects a random sample of 80 customers. The following table shows the number of customers for each brand of refrigerators.

Brand of refrigerators	Numbers of customers	Sample customers
Philip	150	150/480*80=25
Toshiba	90	90/480*80=15
Sharp	120	20
JVC	40	7
Pansonic	80	13
Total	480	80

(i) State the population of interest.

Customers who purchased the refrigerators from the outlet for the past 12 months.

(ii) State the variable of interest in the study.

Level of satisfaction among Customers

(iii) What is the sampling method used by the manage?

Stratified Random Sampling

(iv) Calculate the number of customers selected as samples for each brand.

(v) Describe how to select the customers for the brand using systematic sampling.

K=N/n=480/80=6^{th}

First sample ------1 to 6

Fist sampe: A or 2

Sample numbers: 2, 8,14,20, 26,32,-----

Sample numbers: A, A+6, A+12, A+18,----------, A+474.


Thank you
for your
attention